STATISTICS IN GENETICS

A.W. van der Vaart

Fall 2006, Preliminary Version, corrected and extended Fall 2011 (version 20/7/2015)
Warning: some parts more correct than others
Reading and believing at own risk

CONTENTS

1. Segregation . . . 1
   1.1. Biology . . . 1
   1.2. Mendel's First Law . . . 5
   1.3. Genetic Map Distance . . . 9
   1.4. Inheritance Indicators . . . 19
2. Dynamics of Infinite Populations . . . 23
   2.1. Mating . . . 24
   2.2. Hardy-Weinberg Equilibrium . . . 24
   2.3. Linkage Equilibrium . . . 29
   2.4. Full Equilibrium . . . 31
   2.5. Population Structure . . . 32
   2.6. Viability Selection . . . 33
   2.7. Fertility Selection . . . 49
   2.8. Assortative Mating . . . 50
   2.9. Mutation . . . 52
   2.10. Inbreeding . . . 52
3. Pedigree Likelihoods . . . 55
   3.1. Pedigrees . . . 55
   3.2. Fully Informative Meioses . . . 57
   3.3. Pedigree Likelihoods . . . 60
   3.4. Parametric Linkage Analysis . . . 66
   3.5. Counselling . . . 68
   3.6. Inheritance Vectors . . . 69
   3.7. Elston-Stewart Algorithm . . . 71
   3.8. Lander-Green Algorithm . . . 74
4. Identity by Descent . . . 79
   4.1. Identity by Descent and by State . . . 79
   4.2. Incomplete Data . . . 81
   4.3. Distribution of IBD-indicators . . . 82
   4.4. Conditional Distributions . . . 87
5. Nonparametric Linkage Analysis . . . 90
   5.1. Nuclear Families . . . 90
   5.2. Multiple Testing . . . 94
   5.3. General Pedigrees . . . 96
   5.4. Power of the NPL Test . . . 98
   5.5. Holmans' Triangle . . . 101
6. Genetic Variation . . . 107
   6.1. Variance . . . 107
   6.2. Covariance . . . 114
7. Heritability . . . 126
   7.1. Environmental Influences . . . 126
   7.2. Heritability . . . 127
   7.3. Biometrical Analysis . . . 128

7.4. Regression to the Mean 131

7.5. Prevalence 132

8. Quantitative Trait Loci 134

8.1. Haseman-Elston Regression 135

8.2. Covariance Regression 136

8.3. Copulas 143

8.4. Frailty Models 145

9. Association Analysis 147

9.1. Association and Linkage Disequilibrium 148

9.2. Case-Control Tests 156

10. Combined Linkage and Association Analysis 170

10.1. Transmission Disequilibrium Test 170

10.2. Sibship Transmission Disequilibrium Test 177

11. Coalescents 179

11.1. Wright-Fisher Model 179

11.2. Robustness 188

11.3. Varying Population Size 191

11.4. Diploid Populations 193

11.5. Mutation 194

11.6. Recombination 198

12. Random Drift in Population Dynamics 205

13. Phylogenetic Trees 206

14. Statistics and Probability 207

14.1. Contingency Tables 207

14.2. Likelihood Ratio Statistic 217

14.3. Score Statistic 224

14.4. Multivariate Normal Distribution 229

14.5. Logistic Regression 233

14.6. Variance Decompositions 237

14.7. EM-Algorithm 244

14.8. Hidden Markov Models 246

14.9. Importance Sampling 250

14.10. MCMC Methods 252

14.11. Gaussian Processes 256

14.12. Renewal Processes 257

14.13. Markov Processes 261

14.14. Multiple Testing 266

EXAM

The Fall 2008 exam comprises the material in Chapters 1–10 and 14 that is not marked by asterisks *. The material in Chapter 14 is background material for the main text, important only when quoted, EXCEPT the sections on Variance decompositions, the EM-algorithm, and Hidden Markov models, which are important parts of the course. Do know what a score test, likelihood ratio test, chisquare test etc. are, and how they are carried out.

LITERATURE

[1] Almgren, P., Bendahl, P.-O., Negtsson, H., Hossjer, O. and Perfekt, R., (2004). *Statistics in Genetics.* Lecture notes, Lund University.

[2] Lange, K., (2002). *Mathematical and Statistical Methods for Genetic Analysis, 2nd Edition.* Springer Verlag.

[3] Sham, P., (1997). *Statistics in Human Genetics.* Arnold Publishers.

[4] Thompson, E., (2000). *Statistical Inference from Genetic Data on Pedigrees.* Institute of Mathematical Statistics.

1
Segregation

This chapter first introduces a (minimal) amount of information on genetic biology, and next discusses stochastic models for the process of meiosis.

The biological background discussed in this chapter applies to "most" living organisms, including plants. However, we are particularly interested in human genetics and it will be understood that the discussion refers to humans, or other organisms with the same type of sexual reproduction.

1.1 Biology

The genetic code of an organism is called its *genome* and can be envisioned as a long string of "letters". Physically this string corresponds to a set of *DNA-molecules*, which are present (and identical) in every cell of the body. The genome of an individual is formed at conception and remains the same throughout life, apart from possible mutations and other aberrations during cell division.

The genome of a human is divided over 46 DNA-molecules, called *chromosomes.* These form 23 pairs, 22 of which are called *autosomes*, the remaining pair being the *sex chromosomes.* (See Figure 1.1.) The two chromosomes within a pair are called *homologous.* The sex chromosomes of males are coded XY and are very different; those of females are coded YY. One chromosome of each pair originates from the father, and the other one from the mother. We shall usually assume that the paternal and maternal origins of the chromosomes are not important for their function.

Chromosomes received their names at the end of the 19th century from the fact that during cell division they can be observed under the microscope as elongated molecules that show coloured bands after staining. (See Figure 1.2.) Also visible in every chromosome is a special location somewhere in the middle, called *centromere*, which plays a role in the cell division process. The two pieces of chromosome extending on either side of the centromere are known as the p-arm and q-arm, and

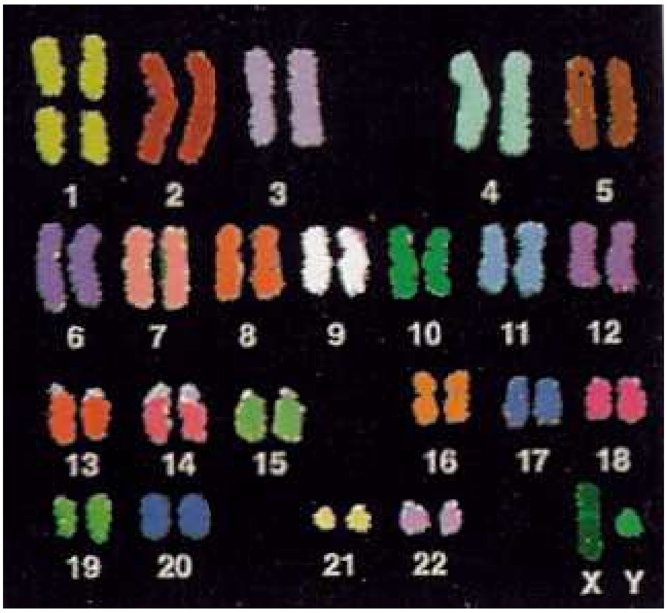

Figure 1.1. The 23 pairs of human chromosomes (of a male) neatly coloured and arranged.

loci on chromosomes are still referred to by codes such as “9q8” (meaning band 8 on the q-arm of chromosome 9). The endpoints are called *telomeres*.

The chemical structure of DNA was discovered in 1959 by Watson and Crick. DNA consists of two chains of *nucleotides* arranged in a double-helix structure. There are four of such nucleotides: Adenine, Citosine, Guanine and Thymine, and it is the first letters A, C, G, T of their names that are used to describe the genetic code. The two chains of nucleotides in DNA carry “complementary letters”, always pairing Adenine to Thymine and Citosine to Guanine, thus forming *base pairs*. Thus a chromosome can be represented by a single string of letters A, C, T, G. The human genome has about 3×10^9 base pairs.

Figure 1.3 gives five views of chromosomes, zooming out from left to right. The left panel gives a schematic view of the spatial chemical structure of the DNA-molecule. The spiralling bands are formed by the nucleotides and are connected by “hydrogen bonds”. The fourth panel shows a pair of chromosomes attached to each other at a centromere. Chromosomes are very long molecules, and within a cell they are normally coiled up in tight bundles. Their spatial structure is influenced by their environment (e.g. surrounding molecules, temperature), and is very important to their chemical behaviour.

The genome can be viewed as a code that is read off by other molecules, which next start the chain of biological processes that is the living organism. Actually only a small part of the DNA code appears to have biological relevance, most of it being *junk-DNA*. The most relevant part are relatively short sequences of letters, called genes, that are spread across the genome. By definition a *gene* is a subsequence of

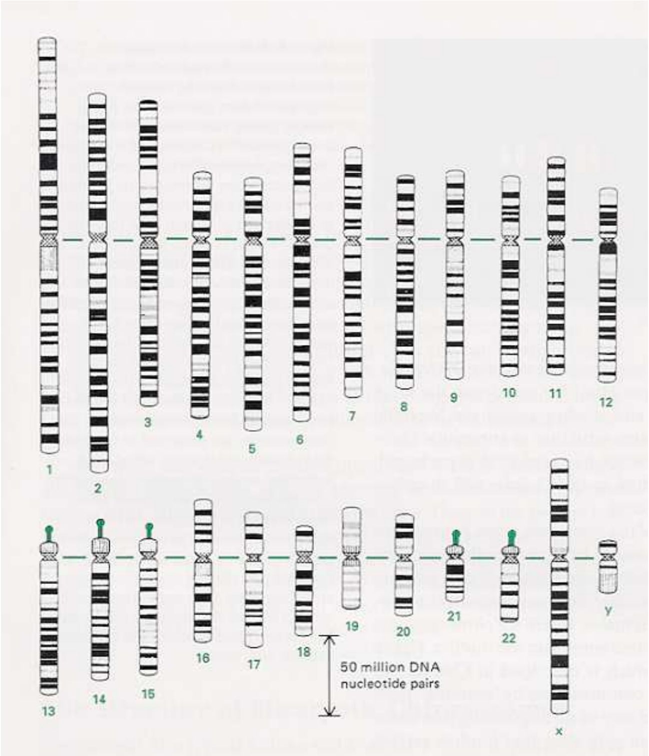

Figure 1.2. One representative of the 23 pairs of human chromosomes aligned on their centromere, showing their relative sizes and the bands that give them their names.

the genome that is translated into a *protein.* A protein is a molecule that consists of a concatenation of *amino-acids.* According to the *central dogma of cell biology* to become active a part of DNA is first *transcribed* into *RNA* and next *translated* into a protein. RNA is essentially a complementary copy (C becomes G, A becomes T, and vice versa) of a part of DNA that contains a gene, where important or *coding* parts, called *exons*, are transcribed, and noncoding parts, called *introns*, are left out. In turn RNA is translated into a *protein*, in a mechanistic way, where each triplet of letters (*codon*) codes for a particular amino-acid. Because there are $4^3 = 64$ possible strings of three nucleotides and only 20?? amino-acids, multiple triplets code for the same protein.

Thus a subsequence of the genome is a *gene* if it codes for some protein. A gene may consist of as many as millions of base pairs, but a typical gene has a length in the order of (tens of) thousands of base pairs. The gene is said to *express its function* through the proteins that it codes for. The processes of transcription and translation are complicated and are influenced by many environmental and genetic factors (promoter, terminator, transcription factors, regulatory elements, methylation, splicing, etc.). The relationship between biological function and the letters coding the gene is therefore far from being one-to-one. However, in (elementary) statistical genetics it is customary to use the genetic code as the explanatory variable, lumping all variations into environmental or “noise” factors.

Because much about the working of a cell is still to be discovered, not all genes are known. However, based on current knowledge and structural analogies it is estimated that the human genome has about 25 000 genes.

The genomes of two individuals are the same to a large extent, and it is even true that the structure of the genomes of different species agrees to a large extent, as the result of a common evolution. It is the small differences that count.

A different variety of a gene is called an *allele*. Here the gene is identified by its location on the genome, its biological function, and its general structure, and the various alleles differ by single or multiple base pairs. In this course we also use the word allele for a segment of a single chromosome that represents a gene, and even for segments that do not correspond to genes.

An individual is called *homozygous* at a locus if the two alleles (the segments of the two chromosomes at that locus) are identical, and *heterozygous* otherwise.

A *locus* refers to a specific part of the genome, which could be a single letter, but is more often a segment of a certain type. A *causal locus* is a locus, typically of a gene, that plays a role in creating or facilitating a disease or another characteristic.

A *marker* is a segment of the genome that is not the same for all individuals, and of which the location is (typically) known. If some observable characteristic (phenotype) is linked to a single genetic locus, then this locus may serve as a marker. Nowadays, markers are typically particular patterns of DNA-letters (RFLPs, VNTRs, Microsatellite polymorphisms, SNPs).

A *haplotype* is a combination of several loci on a single chromosome, often marker loci or genes, not necessarily adjacent.

The *genotype* of an individual can refer to the complete genetic make-up (the set of all pairs of chromosomes), or to a set of specific loci (a pair of alleles or a pair of haplotypes). It is usually opposed to a *phenotype*, which is some observable characteristic of the individual (“blue eyes”, “affection by a disease”, “weight”, etc.). This difference blurs if the genotype itself is observed.

A *single nucleotide polymorphism* (*SNP*, pronounced as “snip”) is a letter on the genome that is not the same for all individuals. “Not the same for all” is interpreted in the sense that at least 1 % of the individuals should have a different letter than the majority. Of the 3×10^9 letters in the human genome only to the order 10^7 letters are SNPs, meaning that more than 99 % of the human genetic code is the same across all humans. The remaining 1 % (the SNPs) occur both in the coding regions (genes) and noncoding regions (junk DNA) of the genome. Two out of three SNPs involve the replacement of Cytosine by Thymine.

1.1.1 Note on Terminology

In the literature the words “gene”, “allele” and “haplotype” are used in different and confusing ways. A *gene* is often viewed as a functional unit sitting somewhere in the genome. In most organisms the autosomal chromosomes occur in pairs and hence these functional units are represented by two physical entities, DNA sequences of a given type. There is no agreement whether to use “gene” for the pair of functionally similar DNA sequences or for each of the two copies. In the latter use each cell contains two genes of each given type. Part of the interest in this course stems from the fact that the DNA sequence for a given gene, even though largely determined, varies across the population and among the two copies of a person in a number of positions. The word *allele* is used for the possible varieties of the DNA sequence, but often also for the physical entity itself, when it becomes equivalent to one of the two uses of the word "gene". In the latter meaning it is also equivalent to a "single-locus haplotype", even though the word *haplotype* is typically reserved for a piece of chromosome containing *multiple* loci of interest. When contemplating this, keep in mind that the exact meaning of the word "locus" can be context-dependent as well. A *locus* is a place on the DNA string. When discussing the action of several genes, each gene is considered to occupy a locus, but when considering a single gene a "locus" may well refer to a single nucleotide. A *marker* is typically a locus at a known position of which the variety in a given individual can be established (easily) using current technology.

That DNA is a double-stranded molecule invites to further confusion, but this fact is actually irrelevant in most of this course.

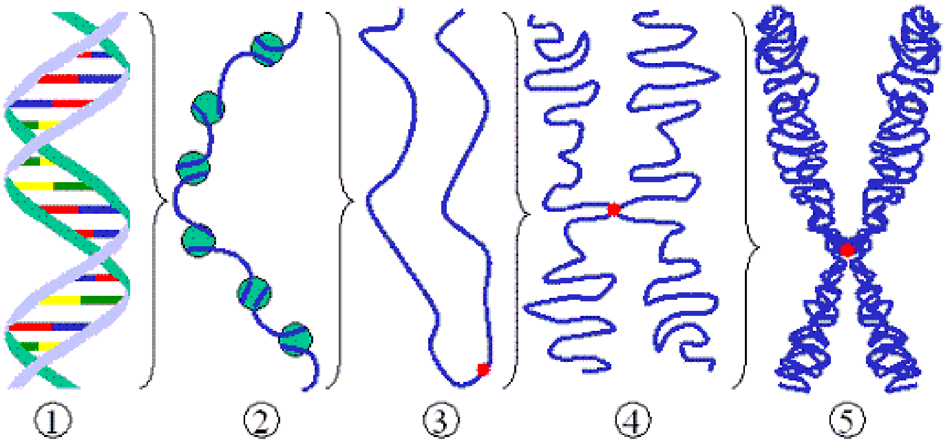

Figure 1.3. Five views of a chromosome. Pictures (4) and (5) show a chromosome together with a copy attached at a "centromere".

1.2 Mendel's First Law

An individual receives the two chromosomes in each pair from his parents, one chromosome from the father and one from the mother. The parents themselves have pairs of chromosomes, of course, but form special cells, called *gametes* (sperm for males and ovum for females), which contain only a single copy of each chromosome.

At conception a sperm cell and ovum unite into a *zygote*, and thus form a cell with two copies of each chromosome. This single cell next goes through many stages of cell division (mitosis) and specialization to form eventually a complete organism.

Thus a parent passes on (or *segregates*) half of his/her genetic material to a child. The single chromosome in a gamete is not simply a copy of one of the two chromosomes of the parent, but consists of segments of both. The biological process by which a parent forms gametes is called *meiosis*, and we defer a discussion to Section 1.3.

Mendel (1822–1884) first studied the *segregation* of genes systematically, and formulated two laws.

Mendel's first law is the *Law of Segregation*: *parents choose the allele they pass on to their offspring at random from their pair of alleles.*

Mendel's second law is the *Law of Assortment*: *segregation is independent for different genes.*

Our formulation using the word "choose" in the first law is of course biologically nonsensical.

Mendel induced his laws from studying the phenotypes resulting from experiments with different varieties of peas, and did not have much insight in the underlying biological processes. The law of segregation is still standing up, but the law of assortment is known to be wrong. Genes that are close together on a single chromosome are not passed on independently, as pieces of chromosome rather than single genes are passed on. On the other hand, genes on different chromosomes are still assumed to segregate independently and hence satisfy also Mendel's second law. In this section we consider only single genes, and hence only Mendel's first law is relevant. In Section 1.3 we consider the segregation of multiple genes.

We shall always assume that the two parents "act" independently. Under Mendel's first law we can then make a segregation table, showing the proportion of offspring given the genotypes of the parents. These segregation ratios are shown in the first two columns of Table 1.1 for a single biallelic gene with alleles A and a. There are 3 possible individuals (AA, Aa and aa) and hence $3 \times 3 = 9$ possible ordered pairs of parents ("mating pairs"). As long as we do not consider the sex chromosomes, we could consider the parents as interchangeable. This is the reason that the first column of the table shows only the 6 different *unordered* pairs of parents. Columns 2–4 show the probabilities of a child having genotype AA, Aa or aa given the parent pair, computed according to Mendel's first law.

The remaining columns of the table show the probabilities of *phenotypes* corresponding to the genotypes under three possible assumptions: dominance, codominance or recession of the allele A. The underlying assumption is that the gene under consideration (with possible genotypes AA, Aa and aa) is the sole determinant of the observable characteristic. The allele A is called *dominant* if the genotypes AA and Aa give rise to the same phenotype, marked "A" (of "affected") in the table, versus a different phenotype, marked "U" (of "unaffected") in the table, corresponding to the genotype aa. The allele A is called *recessive* if the genotypes Aa and aa give rise to the same phenotype, marked "U" in the table, versus a different phenotype corresponding to genotype AA. The probabilities of the phenotypes in these cases are given in the columns marked "dom" and "rec", and can simply be obtained by adding the appropriate columns of genotypic probabilities together. The remaining case is that of *codominance*, in which the three different genotypes give rise to three different phenotypes, marked "1, 2, 3" in the table. The corresponding columns of the table are exact copies of the genotypic columns.

mating pair	offspring			dom		codom			rec	
	AA	Aa	aa	A	U	1	2	3	A	U
$AA \times AA$	1	–	–	1	–	1	–	–	1	–
$AA \times Aa$	$\frac{1}{2}$	$\frac{1}{2}$	–	1	–	$\frac{1}{2}$	$\frac{1}{2}$	–	$\frac{1}{2}$	$\frac{1}{2}$
$AA \times aa$	–	1	–	1	–	–	1	–	–	1
$Aa \times Aa$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$	–	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{4}$
$Aa \times aa$	–	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	–	$\frac{1}{2}$	$\frac{1}{2}$	–	1
$aa \times aa$	–	–	1	–	1	–	–	1	–	1

Table 1.1. Six possible genotypes of unordered pairs of parents, the conditional distribution of the genotypes of their offspring (columns 2–4), and their phenotypes under full penetrance with dominance (columns 5–6), codominance (columns 7–9) and recession (columns 10–11).

For many genotypes the categories of dominance, recession and codominance are too simplistic. If A is a disease gene, then some carriers of the genotype AA may not be affected (*incomplete penetrance*), whereas some carriers of genotype aa may be affected (*phenocopies*). It is then necessary to express the relationship between genotype and phenotype in probabilities, called *penetrances*. The simple situations considered in Table 1.1 correspond to "full penetrance without phenocopies".

Besides, many diseases are dependent on multiple genes, which may have many alleles, and there may be environmental influences next to genetic determinants.

1.1 Example (Blood types). The definitions of dominant and recessive alleles extend to the situation of more than two possible alleles. For example the ABO locus (on chromosome 9q34) is responsible for the 4 different blood phenotypes. The locus has 3 possible alleles: A, B, O, yielding 6 unordered genotypes. Allele O is recessive relative to both A and B, whereas A and B are codominant, as shown in Table 1.2. □

genotype	phenotype
OO	O
AA, AO	A
BB, BO	B
AB	AB

Table 1.2. Genotypes at the ABO locus and corresponding phenotypes.

* 1.2.1 Testing Segregation Propertions

We can test the validity of the Table 1.1 (and hence the recessive, dominant or codominant nature of a single locus model) by several procedures. The general idea is to sample a certain type of individual (mating pair and/or offspring) based on their phenotype and next see if their relatives occur in the proportions as predicted by the table under the various sets of hypotheses.

As an example we shall assume that the allele A is rare, so that the frequency of the genotype AA can be assumed negligible relative to the (unordered) genotype Aa.

(i) Suppose that A is dominant, and we take a sample of n couples consisting of an affected and a healthy parent. Because A is rare, almost all of these couples must be $Aa \times aa$, and hence their offspring should be affected or normal each with probability $\frac{1}{2}$. The total number N of affected offspring is binomially distributed with parameters n and p. We can verify the validity of our assumptions by testing the null hypothesis $H_0: p = \frac{1}{2}$.

(ii) If A is codominant, then we can identify individuals with Aa genotypes from their observed phenotypes, and can take a random sample of $Aa \times Aa$-couples. The total number of offspring (N_1, N_2, N_3) of the three possible types in the sample is multinomially distributed. We test the null hypothesis that the success probabilities are $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.

(iii) Suppose that A is recessive, and we take a random sample of unaffected parents who have at least one affected child. The parents are certain to have genotypes $Aa \times Aa$. Under our sampling scheme the number N^i of affected children in the ith family is distributed as a binomial variable with parameters s^i (the family size) and p conditioned to be positive, i.e.

$$P(N^i = n) = \frac{\binom{s^i}{n} p^n (1-p)^{s^i - n}}{1 - (1-p)^{s^i}}, \qquad n = 1, \ldots, s^i.$$

We test the null hypothesis $H_0: p = \frac{1}{4}$. Computation of the maximum likelihood estimate can be carried out by the EM-algorithm or Fisher scoring.

(iv) Suppose again that A is recessive, and we take a random sample of affected children. Because A is rare, we may assume that the parents of these children are all $Aa \times Aa$. We collect the children into families (groups of childeren with the same parents), and determine for each family the number B of sampled (and hence affected) children and the total number N of affected children in the family. We model N as a binomial variable with parameters the family size s and p, and model B given N as binomial with parameters N and π, where π is the “ascertainment” probability. We observe N and B only if $B > 0$. Under these assumptions

$$\begin{aligned} P(N = n, B = b \mid B > 0) &= \frac{P(B = b \mid N = n) P(N = n)}{Pr(B > 0)} \\ &= \frac{\binom{n}{b} \pi^b (1-\pi)^{n-b} \binom{s}{n} p^n (1-p)^{s-n}}{1 - (1 - \pi p)^s}. \end{aligned}$$

We can estimate the pair (p, π) by maximum likelihood. We test the null hypothesis $H_0 \colon p = \frac{1}{4}$.

1.3 Genetic Map Distance

Mendel's second law is the *Law of Assortment*: *segregation is independent for different loci.* This law is false: genes that are on the same chromosome (called *syntenic* versus *nonsyntenic*) are not passed on independently. To see how they are passed on we need to study the process of the formation of gametes (sperm and egg cells). This biological process is called *meiosis*, involves several steps, and is not the same for every living organism. The following very schematic description of the process of meiosis in humans is sufficient for our purpose.

The end-product of meiosis is a gamete (egg or sperm cell) that contains a single copy (haplotype) of each chromosome. Offspring is then formed by uniting egg and sperm cells of two parents, thus forming cells with pairs of chromosomes.

Gametes, cells with a single chromosome, are formed from *germ cells* with two chromosomes. The first step of meiosis actually goes in the "wrong" direction: each of the two chromosomes within a cell is duplicated, giving four chromosomes, called *chromatids*. The chromatids are two pairs of identical chromosomes, called "sister pairs". These four strands of DNA next become attached to each other at certain loci (the same locus on each chromosome), forming socalled *chiasmata*. Subsequently the four strands break apart again, where at each chiasma different strands of the four original strands may remain bound together, creating *crossovers*. Thus the resulting four new strands are reconstituted of pieces of the original four chromatids.

If the two sister pairs are denoted S, S and S', S', only chiasmata between an S and an S' are counted as true chiasmata in the following, and also only those that after breaking apart involve an S and an S' on each side of the chiasma. See Figures 1.4 and 1.5 for illustration.

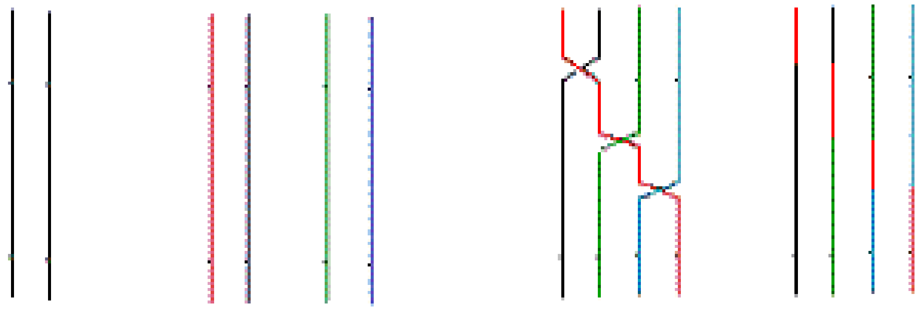

Figure 1.5. Schematic view of meiosis, showing the pair of chromosomes of a single parent on the left, which duplicates and combine into four chromatids on the right. The parent segregates a randomly chosen chromatid. The second panel shows the two pairs of sister chromatids: red and black are identical and so are green and blue. Crossovers within these pairs (e.g. black to red) do not count as true crossovers.

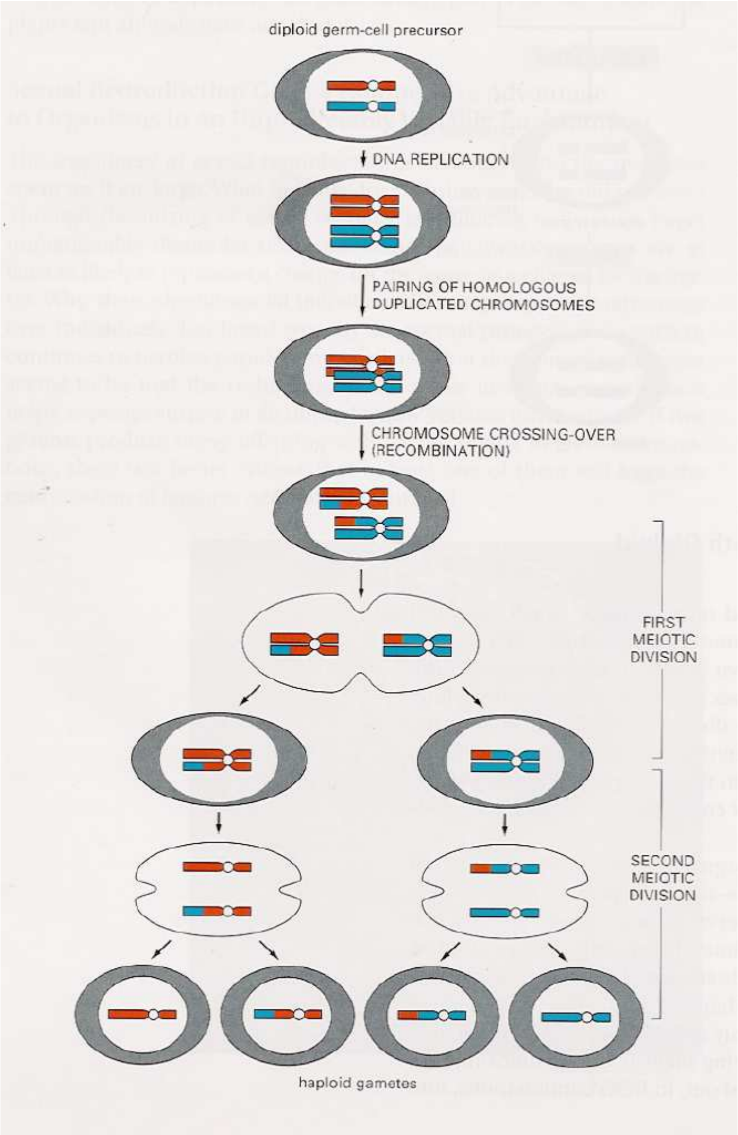
diploid germ-cell precursor
DNA REPLICATION
PAIRING OF HOMOLOGOUS DUPLICATED CHROMOSOMES
CHROMOSOME CROSSING-OVER (RECOMBINATION)
FIRST MEIOTIC DIVISION
SECOND MEIOTIC DIVISION
haploid gametes

Figure 1.4. Realistic view of meiosis.

If we fix two loci and a single chromosome resulting from a meiosis, then we say that there is a *recombination* between the loci if the chromosome at the loci results from different sister pairs S and S'. This is equivalent to there being an odd number of crossovers between the two loci, i.e. the chromosome having been involved in an odd number of chiasmata between the two loci. There may have been other chiasmata between the two loci in which the chosen chromosome has not been involved, as the chiasmata refer to the set of four chromatids. A given chromosome resulting from a meiosis is typically on average involved in half the chiasma. The probability of there being a recombination between two loci of the chromosome of a randomly chosen gamete is known as the *recombination fraction*.

Warning. The word "between" in "recombination between two loci" may lead to misunderstanding. There being recombination between two loci or not depends only on the chromosome (or chromatid) at the two loci, not on what happens at intermediate loci. In particular, if there is no recombination between two given loci, then there may well be recombination between two loci that are in the interval between these given loci.

A stochastic model for the process of meiosis may consist of two parts:

(i) A stochastic process determining the locations of the chiasmata in the four chromatids.

(ii) A stochastic process indicating for each chiasma which two of the four chromatids take part in the chiasma.

The model in (i) is a *point process.* The model in (ii) needs to pick for each chiasma one chomatid from each of the two sister pairs (S,S) and (S',S').

An almost universally accepted model for (ii) is the model of *no chromatid interference (NCI)*, which says that the sisters S and S' are chosen at random and independently from the pairs of sister, for each chiasma, independently across the chiasmata and independently from their placements (i).

The most popular model for the placements of the chiasmata (i) is the Poisson process. Because this tends to give a relatively crude fit to reality, several other models have been suggested. We shall always adopt NCI for (ii), but discuss some alternatives to the Poisson model below. All models for the placement of the chiasmata view the chromatids as lines without structure; in particular they do not refer to the DNA-sequence.

The assumption of NCI readily leads to *Mather's formula.* Fix two loci and consider the experiment of picking at random one of the four strands resulting from a meiosis. Mather's formula concerns the probability of recombination between the two loci.

1.2 Theorem (Mather's formula). *Under the assumption of no chromatid interference, the recombination fraction θ between two given loci satisfies $\theta = \frac{1}{2}(1-p_0)$ for p_0 the probability that there are no chiasmata between the two loci.*

Proof. Let N denote the number of chiasmata between the loci. Under NCI the two chromatids involved in a given chiasma can be considered to be formed by choosing at random a sister chromatid from each of the pairs S,S and S',S'. This includes the chromosome we choose at random from the four strands formed after meiosis (see the description preceding the lemma) with probability $\frac{1}{2}$. Under NCI the chromatids involved in different chiasmata are chosen independently across chiasma. It follows that given $N=n$ the number K of chiasma in which the chosen chromosome is involved is binomially distributed with parameters n and $\frac{1}{2}$. Recombination between the two loci takes place if and only if K is odd. If $n=0$, then $K=0$ and recombination is impossible. If $n > 0$, then

$$P\bigl(K \in \{1,3,5,\ldots\}\,|\,N = n\bigr) = \sum_{k\in\{1,3,5,\ldots\}} \binom{n}{k} (\tfrac{1}{2})^n = \tfrac{1}{2}.$$

The last equality follows easily by recursion, by conditioning the probability that in n fair trials we have an odd number of successes on the event that the first $n-1$ trials produced an odd or even number of successes.

The unconditional probability that K is odd is obtained by multiplying the preceding display by $P(N = n)$ and summing over $n \geq 1$. This is equal to $\frac{1}{2}P(N \geq 1) = \frac{1}{2}(1 - p_0)$. ∎

A consequence of Mather's formula is that the recombination fraction is contained in the interval $[0, \frac{1}{2}]$. If the loci are very close together, then the probability of no chiasmata between them is close to 1 and the recombination fraction is close to $\frac{1}{2}(1-1) = 0$. For distant loci the probability of no chiasmata is close to 0 and the recombination fraction is close to $\frac{1}{2}(1-0) = \frac{1}{2}$. Loci at recombination fraction $1/2$ are called *unlinked.*

Mather's formula can be generalized to the the occurrence of recombination in a collection of intervals. The joint distribution of recombinations can be characterized in terms of the "avoidance probabilities" of the chiasmata process. Fix $k+1$ loci ordered along a chromosome, forming k intervals, and let $R_1, \ldots, R_k$ indicate the occurrence of crossovers between the endpoints of these intervals in a randomly chosen chromatid: $R_j = 1$ if there is a crossover between the endpoints of the jth interval and $R_j = 0$ otherwise. Let $N_1, \ldots, N_k$ denote the numbers of chiasmata in the k intervals in the set of four chromatids.

1.3 Theorem. *Under the assumption of no chromatid interference, for any vector* $(r_1, \ldots, r_k) \in \{0,1\}^k$,

$$P(R_1 = r_1, \ldots, R_k = r_k) = (\tfrac{1}{2})^k \Bigl(1 + \sum_{\substack{S:S\subset\{1,\ldots,k\}\\ S\neq\emptyset}} (-1)^{\sum_{j\in S} r_j} P(N_j = 0\ \forall j \in S)\Bigr).$$

Proof. Let $K_1, \ldots, K_k$ be the numbers of chiasmata in the consecutive intervals in which the chromatid is involved. Under NCI given $N_1, \ldots, N_k$ these variables are independent and K_j has a binomial distribution with parameters N_j and $\frac{1}{2}$. A crossover occurs ($R_j = 1$) if and only if K_j is odd. As in the proof of Mather's formula it follows that $P(K_j \text{ is odd}|\, N_j)$ is $\frac{1}{2}$ if $N_j > 0$; it is clearly 0 if $N_j = 0$. In other words $P(R_j = 1|\, N_j) = \frac{1}{2}(1 - 1_{N_j=0})$, which implies that $P(R_j = 0|\, N_j) = \frac{1}{2}(1 + 1_{N_j=0})$. In view of the conditional independence of the R_j, this implies that

$$\begin{aligned} P(R_1 = r_1, \ldots, R_k = r_k) &= \mathrm{E}P(R_1 = r_1, \ldots, R_k = r_k|\, N_1, \ldots, N_k) \\ &= \mathrm{E} \prod_{j:r_j=1} \tfrac{1}{2}(1 - 1_{N_j=0}) \prod_{j:r_j=0} \tfrac{1}{2}(1 + 1_{N_j=0}). \end{aligned}$$

The right side can be rewritten as the right side of the theorem. ∎

For $k = 1$ the assertion of the theorem reduces to Mather's formula. For $k = 2$ it gives the identities

$$\begin{aligned}
4P(R_1 = 1, R_2 = 1) &= 1 + P(N_1 = 0, N_2 = 0) - P(N_1 = 0) - P(N_2 = 0),\\
4P(R_1 = 1, R_2 = 0) &= 1 - P(N_1 = 0, N_2 = 0) - P(N_1 = 0) + P(N_2 = 0),\\
4P(R_1 = 0, R_2 = 1) &= 1 - P(N_1 = 0, N_2 = 0) + P(N_1 = 0) - P(N_2 = 0),\\
4P(R_1 = 0, R_2 = 0) &= 1 + P(N_1 = 0, N_2 = 0) + P(N_1 = 0) + P(N_2 = 0).
\end{aligned}$$

For general k the formula shows how the process of recombinations can be expressed in the avoidance probabilities of the chiasmata process. A general *point process* on $(0, \infty)$ can be described both as an ordered sequence of positive random variables $S_1 < S_2 < \cdots$, giving the points or "events" of the process, and as a set of random variables $\bigl(N(B): B \in \mathcal{B}\bigr)$ giving the numbers of points $N(B) = \#(i: S_i \in B)$ falling in a (Borel) set B. The *avoidance probabilities* are by definition the probabilities $P\bigl(N(B) = 0\bigr)$ that a set set B receives no points. Because it can be shown that the avoidance probabilities determine the complete point process[†], it is not surprising that the recombination probabilities can be expressed in some way in the avoidance probabilities of the chiasmata process. The theorem makes this concrete.

The *genetic map distance* between two loci is defined as the expected number of crossovers between the loci, on a single, randomly chosen chromatid. The unit of genetic map distance is the *Morgan*, with the interpretation that a distance of 1 Morgan means an expected number of 1 crossover in a single, randomly chosen chromatid. The genetic map length of the human male autosomal genome is about 28.5 Morgan and of the human female genome about 43 Morgan. Thus there are somewhat more crossovers in females than in males, and on the average there are are about 1-2 crossovers per chromosome.

Because expectations are additive, genetic map distance is a linear distance, like the distance on the real line: the distance between loci A and C for loci A, B, C that are physically placed in that order is the sum of the distance between A and B and the distance between B and C. For a formal proof define K_{AB}, K_{BC} and K_{AC} to be the number of crossovers on the segments A–B, B–C and A–C. By definition the genetic map lengths of the three segments are $m_{AB} = \mathrm{E}K_{AB}$, $m_{BC} = \mathrm{E}K_{BC}$ and $m_{AC} = \mathrm{E}K_{AC}$. Additivity: $m_{AC} = m_{AB} + m_{BC}$ follows immediately from the identity $K_{AC} = K_{AB} + K_{BC}$.

The chemical structure of DNA causes that genetic map distance is not linearly related to *physical distance*, measured in base pairs. For instance, *recombination hotspots* are physical areas of the genome where crossovers are more likely to occur. Correspondingly, there exists a linkage map and a physical map of the genome, which do no agree. See Figure 1.6. From a modern perspective physical distance is the more natural scale. The main purpose of genetic map distance appears to be to translate recombination probabilities into a linear distance.

A definition as an "expected number" of course requires a stochastic model (as in (i)–(ii)). (An alternative would be to interpret this "expectation" as "empirical

[†] E.g. Van Lieshout, Markov Point Processes and Their Applications, 2000, Theorem 1.2

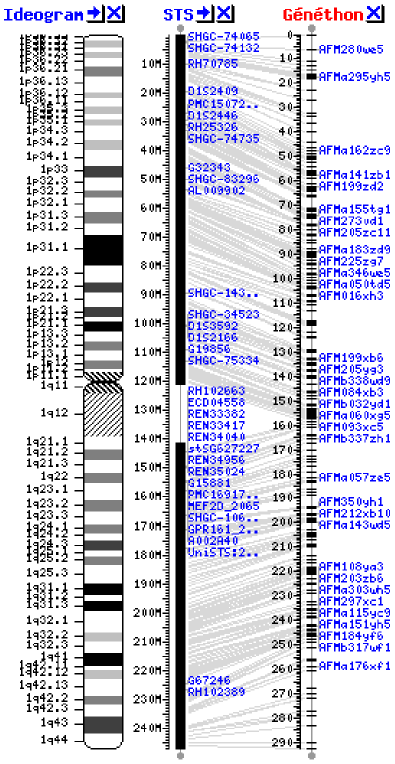

Figure 1.6. Ideogram of chromosome 1 (left), a physical map (middle), and a genetic map (right) with connections between the physical and genetic map shown by lines crossing the displays. (Source NCBI map viewer, Homo Sapiens, Build 36, `http://www.ncbi.nlm.nih.gov/mapview`). The ideogram shows on the left the classical method of addressing genomic positions in terms of the p- and q-arms and numbered coloured bands. The STS- and Généthon-maps are given together with rulers showing position in terms of base pairs (0–240 000 000 bp) and centi-Morgan (0–290 cM), respectively. Corresponding positions on the rulers are connected by a line.

average".) The most common model for the locations of the chiasmata is the Poisson process. We may think of this as started at one end of the chromatids; or we may think of this as generated in two steps: first determine a total number of chiasmata for the chromatids according to the Poisson distribution and next distribute this number of chiasmata randomly uniformly on the chromatids. The Poisson process must have intensity 2 per Morgan, meaning that the expected number of chiasmata per Morgan is 2. Since each chromatid is involved on the average in $\frac{1}{2}$ the chiasmata, this gives the desired expected number of 1 crossover per Morgan in a single chromatid.

Under the Poisson model the probability of no chiasmata between two loci that are m Morgans apart is equal to e^{-2m}. By Mather's formula (valid under NCI) this gives a recombination fraction of

$$\theta = \tfrac{1}{2}(1 - e^{-2m}).$$

The map $m \mapsto \theta(m) = \frac{1}{2}(1 - e^{-2m})$ is called the *Haldane map function*.

Because an independent binomial thinning of a Poisson process is again a Poisson process, under NCI and Haldane's model the process of crossovers on a single chromatid (i.e. in a segregated chromosome) is a Poisson process with intensity 1 per Morgan. The intensity 2 per Morgan is halved, because each chromatid is involved in a given chiasma with probability half.

Because statistical inference is often phrased in terms of recombination fractions, it is useful to connect recombination fractions and map distance in a simple way. In general a *map function* maps the genetic distance into the recombination fraction between loci. The Haldane map function is the most commonly used map function, but several other map functions have been suggested. For instance,

$$\theta(m) = \begin{cases} \frac{1}{2}\tanh(2m), & \textit{Kosambi}, \\ \frac{1}{2}\Bigl(1-\Bigl(1-\frac{m}{L}\Bigr)e^{-m(2L-1)/L}\Bigr), & \textit{Sturt}. \end{cases}$$

The Sturt function tries to correct the fact that in the Haldane model there is a positive probability of no chiasmata in a chromosome. In the Sturt model L is the length of the chomosome in Morgans and the process of chiasmata consists of adding to a Poisson process of intensity $(2L-1)/L$ a single chiasma placed at a random location on the chromosome independently of the Poisson model.

1.4 EXERCISE. Give a formal derivation of the Sturt map function, using the preceding description.

For the Poisson process model the occurrence of crossovers in disjoint intervals is independent, which is not entirely realistic. Other map functions may be motivated by relaxing this assumption. Given ordered loci A, B, C recombination takes place in the interval *A–C* if and only if recombination takes place in exactly one of the two subintervals *A–B* and *B–C*. Therefore, independence of crossovers occurring in the intervals *A–B* and *B–C* implies that the recombination fractions $\theta_{AC}, \theta_{AB}, \theta_{BC}$ of the three intervals *A–C*, *A–B* and *B–C* satisfy the relationship

$$\theta_{AC} = \theta_{AB}(1-\theta_{BC}) + (1-\theta_{AB})\theta_{BC} = \theta_{AB} + \theta_{BC} - 2\theta_{AB}\theta_{BC}.$$

Other map functions may be motivated by replacing the 2 in the equation on the right side by a smaller number $2c$ for $0 \le c \le 1$. The extreme case $c = 0$ is known as *interference* and corresponds to mutual exclusion of crossovers in the intervals *A–B* and *B–C*. The cases $0 < c < 1$ are known as *coincidence*. If we denote the genetic lengths of the intervals *A–B* and *B–C* by m and d and the map function by θ, then we obtain

$$\theta(m+d) = \theta(m) + \theta(d) - 2c\,\theta(m)\theta(d).$$

A map function must satisfy $\theta(0) = 0$. Recombination fraction and map distance are comparable at small distances if $\theta'(0) = 1$. Assuming that θ is differentiable with $\theta(0) = 0$ and $\theta'(0) = 1$, we readily obtain from the preceding display that $\theta'(m) = 1 - 2c\theta(m)$ and hence

$$\theta(m) = \frac{1}{2c}\bigl(1 - e^{-2cm}\bigr).$$

The case $c = 1$ is the Haldane model. Several other map functions can be motivated by using this formula with c a function that depends on m. For instance, the *Carter-Falconer* and *Felsenstein* models correspond to $c = 8\theta(m)^3$ and $c = K - 2\theta(m)(K - 1)$, respectively.

Such ad-hoc definitions have the difficulty that they may not correspond to any probability model for the chiasmata process. A more satisfying approach is to construct a realistic point process model for the chiasmata process and to derive a map function from this. First we need to give a formal definition of a map function, given a chiasmata process $\bigl(N(B), B \in \mathcal{B}\bigr)$. According to Mather's formula, for any interval B, the quantity $\frac{1}{2}P\bigl(N(B) > 0\bigr)$ is the recombination fraction over the interval B. The idea of a map function is to write this as a function of the genetic length (in Morgan) of the interval B, which is by definition $\frac{1}{2}\mathrm{E}N(B)$. Therefore we define $\theta: [0, \infty) \to [0, \frac{1}{2}]$ to be the *map function* corresponding to the chiasmata process N if, for all (half-open) intervals B,

$$\theta\bigl(\tfrac{1}{2}\mathrm{E}N(B)\bigr) = \tfrac{1}{2}P\bigl(N(B) > 0\bigr). \tag{1.5}$$

The existence of such a map function requires that the relationship between the expected values $\mathrm{E}N(B)$ and probabilities $P\bigl(N(B) = 0\bigr)$ be one-to-one, if B ranges over the collection of intervals. This is not true for every chiasmata process N??, but is true for the examples considered below.

If we would strengthen the requirement, and demand that (1.5) be valid for every finite union of disjoint (half-open) intervals, then a map function θ exists only for the case of count-location processes, described in Example 1.8.[‡] This is unfortunate, because according to Theorem 1.3 the joint distribution of recombinations in a set of intervals $B_1, \ldots, B_k$ can be expressed in the probabilities of having no chiasmata in the unions $\cup_{j\in S} B_j$ of subsets of these intervals; equivalently in the probabilities $P\bigl(N(\cup_{j\in S} B_j) > 0\bigr)$. It follows that for general chiasmata processes these probabilities cannot be expressed in the map function, but other characteristics of the chiasmata process are involved.

1.6 EXERCISE. Show that the probability of recombination in both of two *adjacent* intervals can be expressed in the map function as $\frac{1}{2}\bigl(\theta(m_1) + \theta(m_2) - \theta(m_1 + m_2)\bigr)$, where m_1 and m_2 are the genetic lengths of the intervals. [hint: Write $2P(R_1 = 1, R_2 = 1) = \theta(\frac{1}{2}\mathrm{E}N_1) + \theta(\frac{1}{2}\mathrm{E}N_2) - \frac{1}{2}\theta\bigl(\frac{1}{2}(\mathrm{E}N_1 + \mathrm{E}N_2)\bigr)$.]

1.7 Example (Poisson process). The Poisson process N with intensity 2 satisfies $\frac{1}{2}\mathrm{E}N(B) = \lambda(B)$ and $P\bigl(N(B) > 0\bigr) = 1 - e^{-2\lambda(B)}$. Hence the Haldane map function is indeed the map function of this process according to the preceding definition. □

[‡] See Evans et al. (1993).

1.8 Example (Count-location process). Given a probability distributution (p_n) on $\mathbb{Z}^+$ and a probability distribution F on an interval I, let a point process N be defined structurally by first deciding on the total number of points $N(I)$ by a draw from (p_n) and next distributing these $N(I)$ points as the order statistics of a random sample of size $N(I)$ from F.

Given $N(I) = n$ the number of points $N(B)$ in a set B is distributed as the random variable $\sum_{i=1}^n 1_{X_i \in B}$, for $X_1, \ldots, X_n$ a random sample from F. It follows that $\mathrm{E}\bigl(N(B) |\, N(I) = n\bigr) = nF(B)$ and hence $\mathrm{E}N(B) = \mu F(B)$, for $\mu = \mathrm{E}N(I)$.

Given $N(I) = n$ no point falls in B if and only if all n generated points end up outside B, which happens with probability $\bigl(1 - F(B)\bigr)^n$. Therefore, by a similar conditioning argument we find that

$$P\bigl(N(B) = 0\bigr) = \sum_n p_n \bigl(1 - F(B)\bigr)^n = M\bigl(1 - F(B)\bigr),$$

for $M(s) = \mathrm{E}s^{N(I)}$ the moment generating function of the variable $N(I)$. It follows that (1.5) holds with map function $\theta(m) = \frac{1}{2}M(1 - 2m/\mu)$. Equation (1.5) is true even for every Borel set B.

The Poisson process is the special case that the total number of points $N(I)$ possesses a Poisson distribution and F is the uniform distribution on I.

In this model, given the number of points that fall inside in a set B, the locations of these points are as the order statistics of a random sample from the restriction of F to B and they are stochastically independent from the number and locations of the points outside B. This is not considered realistic as a way of modelling possible interference of the locations of the chiasmata, because one would expect that the occurrence of chiasmata near the boundary of B would have more influence on chiasmata inside B than more distant chiasmata. □

1.9 EXERCISE. Prove the assertion in the preceding paragraph.

1.10 Example (Renewal processes). A *stationary renewal process* on $[0, \infty)$ is defined by points at the locations $E_1, E_1 + E_2, E_1 + E_2 + E_3, \ldots$, where $E_1, E_2, E_3, \ldots$ are independent positive random variables, $E_2, E_3, \ldots$ having a distribution F with finite expectation $\mu = \int_0^\infty x\, dF(x)$ and E_1 having the distribution F_1 with density $(1 - F)/\mu$. The exceptional distribution of E_1 makes the point process stationary in the sense that the shifted process of counts $\bigl(N(B + h) \colon B \in \mathcal{B}\bigr)$ has the same distribution as $\bigl(N(B) \colon B \in \mathcal{B}\bigr)$, for any $h > 0$. (Because a renewal process is normally understood to have $E_1, E_2, \ldots$ i.i.d., the present process is actually a *delayed renewal process*, which has been made stationary by a special choice of the distribution of the first event.)

The fact that the distribution of $N(B + h)$ is independent of the shift h, implies that the mean measure $\mu(B) = \mathrm{E}N(B)$ is shift-invariant, which implies in turn that it must be proportional to the Lebesgue measure. The proportionality constant can be shown to be the inverse μ^{-1} of the expected time between two events. Thus, for an interval $(a, b]$ we have $\mathrm{E}N\bigl((a, b]\bigr) = (b - a)/\mu$. Because the unit of genetic distance is the Morgan, we must have on the average 2 chiasmata per unit, implying that μ must be $\frac{1}{2}$, whence $\mathrm{E}N\bigl((a,b]\bigr) = 2(b-a)$.

There is at least one event in the interval $(0, b-a]$ if and only if the first event E_1 occurs before $b-a$. Together with the stationarity, this shows that $P\bigl(N\bigl((a,b]\bigr) > 0\bigr) = P(E_1 \leq b-a) = F_1(b-a)$. Together, these observations show that a map function exists and is given by $\theta(m) = \frac{1}{2}F_1(m)$.

It can be shown[b] that any function θ on a finite interval $(0, L]$ with $\theta(0) = 0$, $\theta'(0) = 1$, $\theta' \geq 0$, $\theta'' \leq 0$, $\theta(L) < 1/2$ and $\theta'(L) > 0$ arises in this form from some renewal process. From the relation $\theta = \frac{1}{2}F_1$ and the definition of F_1, it is clear that F can then be recovered from θ through its density $f = \theta''$.

The Poisson process is the special case that all E_j possess an exponential distribution with mean 2. A simple extension of the Poisson model that fits the available data reasonably well is to replace the exponential distribution of $E_2, E_3, \ldots$ by a (scaled) chisquare distribution, the exponential distribution being the special case of a chisquare distribution with 2 degrees of freedom. This is known as the *Poisson skip model.* □

* **1.11** EXERCISE. Find $P\bigl(N(B) = 0\bigr)$ for B the union of two disjoint intervals and N a stationary renewal process.

1.12 EXERCISE. Show that any stationary point process permits a map function.

1.3.1 Simplified View of Meiosis

For most purposes it is not necessary to consider the true biological mechanism of meiosis and the following simplistic (but biologically unrealistic) view suffices. We describe it in silly language that we shall employ often in the following. "A parent lines up the two members of a pair of homologous chromosomes, cuts these chromosomes at a number of places, and recombines the pieces into two new chromosomes by gluing the pieces together, alternating the pieces from the two chromosomes (taking one part from the first chromosome, a second part from the other chromosome, a third part from the first chromosome, etc.). The cut points are called *crossovers.* Finally, the parent chooses at random one of the two reconstituted chromosomes and passes this on to the offspring."

If we thus eliminate the duplication of the chromatids, the expected number of chiasmata (which are now identical to crossovers) should be reduced to 1 per Morgan.

With this simplified view we loose the relationship between the chiasmata and crossover processes, which is a random thinning under NCI. Because a random thinning of a Poisson process is a Poisson process, nothing is lost under Haldane's model. A randomly thinned renewal process is also a renewal process, but with

[b] Zhao and Speed (1996), Genetics 142, 1369–1377.

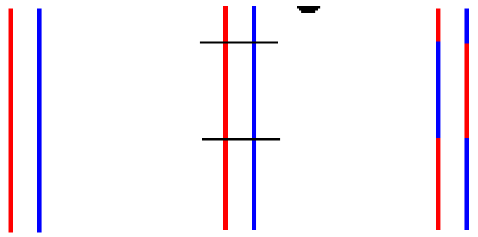

Figure 1.7. Simplified (unrealistic) view of meiosis. The two chromosomes of a single parent on the left cross to produce two mixed chromosomes on the right. The parent segregates a randomly chosen chromosome from the pair on the right.

a renewal distribution of a different shape, making the relationship a bit more complicated.

1.4 Inheritance Indicators

The formation of a child (or zygote) involves two meioses, one paternal and one maternal. In this section we define two processes of *inheritance indicators*, which provide useful notation to describe the crossover processes of the two meioses. First for a given locus u we define two indicators P_u and M_u by

$$P_u = \begin{cases} 0, & \text{if the child's paternal allele is grandpaternal,} \\ 1, & \text{if the child's paternal allele is grandmaternal.} \end{cases}$$

$$M_u = \begin{cases} 0, & \text{if the child's maternal allele is grandpaternal,} \\ 1, & \text{if the child's maternal allele is grandmaternal.} \end{cases}$$

These definitions are visualized in Figure 1.8, which shows a pedigree of two parents and a child. The father is represented by the square and has genotype $(1, 2)$ at the given locus; the mother is the circle with genotype $(3, 4)$; and the child has genotype $(1, 3)$. The genotypes are understood to be ordered by parental origin, with the the paternal allele (the one that is received from the father) written on the left and the maternal allele on the right. In the situation of Figure 1.8 both inheritance indicators P_u and M_u are 0, because the child received the grandpaternal allele (the left one) from both parents.

The inheritance indicators at multiple loci $u_1, \ldots, u_k$, ordered by position on the genome, can be collected together into stochastic processes $P_{u_1}, P_{u_2}, \ldots, P_{u_k}$ and $M_{u_1}, M_{u_2}, \ldots, M_{u_k}$. As the two meioses are assumed independent, these processes are independent. On the other hand, the variables within the two processes are in general dependent. In fact, two given indicators P_{u_i} and P_{u_j} are either equal, $P_{u_i} = P_{u_j}$, or satisfy $P_{u_i} = 1 - P_{u_j}$, where the two possibilities correspond to the nonoccurrence or occurrence of a recombination between loci u_i and u_j in the paternal meiosis. If the loci are very far apart or on different chromosomes, then recombination occurs with probability $\frac{1}{2}$ and the two variables P_{u_i} and P_{u_j} are

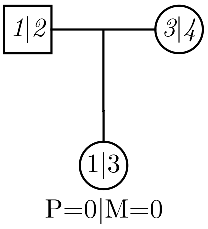

Figure 1.8. Inheritance indicators for a single locus. The two parents have ordered genotypes $(1,2)$ and $(3,4)$, and the child received allele 1 from its father and allele 3 from its mother. Both inheritance indicators are 0.

independent, but if the two loci are linked the two indicators are dependent. The dependence can be expressed in the void probabilities of the chiasmata process, in view of Theorem 1.3. In this section we limit ourselves to the case of the Haldane/Poisson model.

Under the Haldane/Poisson model crossovers occur according to a Poisson process with intensity 1 per unit Morgan. Because the occurrence and locations of events of the Poisson process in disjoint intervals are independent, recombinations across disjoint adjacent intervals are independent and hence the joint distribution of $P=(P_{u_1},P_{u_2},\ldots,P_{u_k})$ can be expressed in the recombination fractions $\theta_1,\ldots,\theta_k$ between the loci, by multiplying the probabilities of recombination or not. This yields the formula

$$P(P=p)=\tfrac{1}{2}\prod_{j=2}^{k}\theta_j^{p_j}(1-\theta_j)^{1-p_j},\qquad p\in\{0,1\}^k.$$

For instance, Table 1.3 gives the joint distribution of $P=(P_{u_1},\ldots,P_{u_k})$ for $k=3$. For simplicity one often takes the distributions of P and M to be the same, although the available evidence suggests to use different values for the recombination fractions for male and female meioses.

In fact, this formula shows that the sequence of variables $P_{u_1},P_{u_2},\ldots,P_{u_k}$ is a discrete time Markov chain (on the state space $\{0,1\}$). A direct way to see this is to note that given $P_{u_1},\ldots,P_{u_j}$ the next indicator $P_{u_{j+1}}$ is equal to P_{u_j} or $1-P_{u_j}$ if there is an even or odd number of crossovers in the interval between loci u_j and u_{j+1}, respectively. The latter event is independent of $P_{u_1},\ldots,P_{u_{j-1}}$, as the latter indicators are completely determined by crossovers to the left of locus u_j. The Markov chain $P_{u_1},P_{u_2},\ldots,P_{u_k}$ is not time-homogeneous. The transition matrix (on the state space $\{0,1\}$) at locus u_j is equal to

$$\begin{pmatrix}1-\theta_j & \theta_j\\ \theta_j & 1-\theta_j\end{pmatrix},\tag{1.13}$$

p	$P(P=p)$
$0,0,0$	$\frac{1}{2}(1-\theta_1)(1-\theta_2)$
$0,0,1$	$\frac{1}{2}(1-\theta_1)\theta_2$
$0,1,0$	$\frac{1}{2}\theta_1(1-\theta_2)$
$0,1,1$	$\frac{1}{2}\theta_1\theta_2$
$1,0,0$	$\frac{1}{2}\theta_1(1-\theta_2)$
$1,0,1$	$\frac{1}{2}\theta_1\theta_2$
$1,1,0$	$\frac{1}{2}(1-\theta_1)\theta_2$
$1,1,1$	$\frac{1}{2}(1-\theta_1)(1-\theta_2)$

Table 1.3. Joint distribution of the inheritance vector $P = (P_{u_1}, P_{u_2}, P_{u_3})$ for three ordered loci u_1–u_2–u_3 under the Haldane model for the chiasmata process. The parameters θ_1 and θ_2 are the recombination fractions between the loci u_1–u_2 and u_2–u_3, respectively.

where θ_j is the recombination fraction for the interval between loci j and $j+1$. The initial distribution, and every other marginal distribution, is binomial with parameters 1 and $\frac{1}{2}$.

The description as a Markov process becomes even more attractive if we think of the inheritance indicators as processes indexed by a locus u ranging over an (idealized) continuous genome. Let $U \subset \mathbb{R}$ be an interval in the real line that models a chromosome, with the ordinary distance $|u_1 - u_2|$ understood as genetic distance in Morgan. The inheritance processes $(P_u: u \in U)$ and $(M_u: u \in U)$, then become continuous time Markov processes on the state space $\{0,1\}$. In fact, as a function of the locus u the process $u \mapsto P_u$ switches between its two possible states 0 and 1 at the locations of crossovers in the meiosis. Under the Haldane/Poisson model these crossovers occur at the events of a Poisson process of intensity 1 (per Morgan, on a single chromatid). If N is this Poisson process, started at one end of the chromosome, then P_u takes the values 0 and 1 either if N_u is even and odd, respectively, or if N_u is odd and even. In the first case $I = M \bmod 2$ and in the second it is $P = N \bmod 2 + 1$. The distribution of the process $u \mapsto P_u$ follows from the following lemma.

1.14 Lemma. *If N is a Poisson process with intensity λ, then the process $[N] = N \bmod 2$ is a continuous time Markov process with transition function*

$$P\bigl([N]_t = 1|\, [N]_s = 0\bigr) = P\bigl([N]_t = 0|\, [N]_s = 1\bigr) = \tfrac{1}{2}(1 - e^{-2\lambda|s-t|}).$$

Proof. For $s < t$ the process $[N]$ changes value across the interval $(s,t]$ if and only if the process N has an odd number of events in this interval. This happens with probability

$$\sum_{k \text{ odd}} e^{-\lambda(t-s)} \frac{\bigl(\lambda(t-s)\bigr)^k}{k!}.$$

This sum can be evaluated as claimed using the equality $e^x - e^{-x} = 2\sum_{k \text{ odd}} x^k/k!$, which is clear from expanding the exponential functions in their power series’.

The Markov property of $[N]$ is a consequence of the fact that the Poisson process has no memory, and that a transition of $[N]$ in the interval $(s,t]$ depends only on the events of N in $(s,t]$. ■

To obtain the distribution of the inheritance processes we choose $\lambda = 1$ in the lemma. The transition probability over an interval of length m in the lemma then becomes $\frac{1}{2}(1 - e^{-2m})$, in which we recognize the Haldane map function.

Markov processes in continuous time are often specified by their generator matrix (see Section 14.13). For the inheritance processes this takes the form

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{1.15}$$

A corresponding schematic view of the process $u \mapsto P_u$ is given in Figure 4.3. The two circles represent the states 0 and 1 and the numbers on the arrows the intensities of transition between the two states.

1.16 Corollary. *Under the Haldane/Poisson model for crossovers the inheritance processes $u \mapsto P_u$ and $u \mapsto M_u$ are independent stationary continuous time Markov processes on the state space $\{0,1\}$ with transition function as given in Lemma 1.14 with $\lambda = 1$ and generator matrix (1.15).*

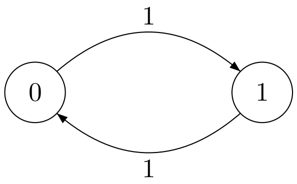

Figure 1.9. The two states and transition intensities of the Markov processes $u \mapsto P_u$ and $u \mapsto M_u$, under the Haldane/Poisson model for crossovers.

2
Dynamics of Infinite Populations

In this chapter we consider the evolution of populations in a discrete-time framework, where an existing population (of parents) is successively replaced by a new population (of children). The populations are identified with a set of possible genotypes and their relative frequencies, and are considered to have infinite size. A children's population can then be described by the probability that an arbitrary child has a certain genotype, a probability that is determined by the likelihoods of the various parent pairs and the laws of meiosis. The laws of meiosis were described in Chapter 1, but may be augmented by allowing mutation.

The simplest model for the formation of parent pairs is the union of independently and randomly chosen parents. This leads to populations that are in Hardy-Weinberg and linkage equilibrium, an assumption that underlies many methods of statistical analysis. We describe this equilibrium in Sections 2.1 to 2.4, which are sufficient background for most of the remaining chapters of the book. In the other sections we consider various types of deviations of random mating, such as selection and assortative mating.

Consideration of infinite rather than finite populations ignores *random drift.* This term is used in genetics to indicate that the relative frequency of a genotype in a finite population of children may deviate from the probability that a random child is of the particular type. A simple model for random drift is to let the frequencies of the genotypes in the next population follow a multinomial vector with N trials and probability vector $(p_g\colon g \in G)$, where p_g is the probability that a child carries genotype g. Under this model the expected values of the relative frequencies in the children's population are equal to $(p_g\colon g \in G)$, but the realized relative frequencies typically will not. Any realized relative frequencies are possible, although in a big population with high probability the realized values will be close to $(p_g\colon g \in G)$.

Models for the randomness of the dynamics of finite populations are discussed in Chapter 12. There also the somewhat artificial structure of separated, nonoverlapping generations is dropped, and evolution is described in continuous time.

2.1 Mating

Consider a sequence of populations of individuals, the $(n+1)$th population consisting of the offspring of the nth population. Identify each individual with a genotype, so that each population is fully described by the vector of relative frequencies of the various genotypes. A prime interest is in the evolution of this vector as a function of generation n.

Assume that the populations are of infinite size and that the $(n+1)$th population arises from the nth by infinitely often and independently creating a single child according to a fixed chance mechanism. The relative frequencies of the genotypes in the $(n+1)$th population are then the probabilities that a single child possesses the various genotypes.

The mechanism to create a child consists of choosing a pair of parents, followed by two meioses, a paternal and a maternal one, which produce two gametes that unite to a zygote. The meioses are assumed to follow the probability models described in Chapter 1, apart from the possible addition of mutation. In most of the chapter we do not consider mutation, and therefore agree to assume its absence, unless stated otherwise. Then the dynamics of the sequence of populations are fixed once it is determined which pairs of parents and with what probabilities produce offspring.

The simplest assumption is *random mating* without selection. This entails that the two parents are independently chosen at random from the population. Here one could imagine separated populations of mothers and fathers, but for simplicity we make this distinction only when considering loci on the sex-chromosomes.

Even though random mating underlies most studies in quantitative genetics, it may fail for many reasons. Under *assortative mating* individuals choose their mates based on certain phenotypes. Given *population structure* individuals may mate within subpopulations, with possible migrations between the subpopulations. By *selection* certain potential parent pairs may have less chance of being formed or of producing offspring. We consider these deviations after describing the basics of plain random mating.

2.2 Hardy-Weinberg Equilibrium

A population is said to be in *Hardy-Weinberg equilibrium* (HW) at a given locus if the two alleles at this locus of a randomly chosen person from the population are stochastically independent and identically distributed. More precisely, if there are k possible alleles $A_1, \ldots, A_k$ at the locus, which occur with relative frequencies $p_1, \ldots, p_k$ in the population, then the ordered pair of alleles at the given locus of a randomly chosen person is (A_i, A_j) with probability $p_i p_j$.

Instead of *ordered genotypes* (A_i, A_j), we can also consider unordered genotypes, which are sets $\{A_i, A_j\}$ of two alleles. This would introduce factors 2 in the Hardy-Weinberg frequencies. If $A_i \neq A_j$, then the unordered genotype

$\{A_i, A_j\}$ results from both A_iA_j and A_jA_i and hence has Hardy-Weinberg frequency $p_ip_j + p_jp_i = 2p_ip_j$. On the other hand, the unordered genotype $\{A_i, A_i\}$ corresponds uniquely to the ordered genotype A_iA_i and has Hardy-Weinberg frequency $p_ip_i = p_i^2$. Generally speaking, ordered genotypes are conceptually simpler, but unordered genotypes are sometimes attractive, because there are fewer of them. Moreover, even though we can always conceptually order the genotypes, for instance by parental origin (with A_i segregated from the father and A_j by the mother), typically only unordered genotypes are observable.

It is a common assumption in statistical inference that a population is in Hardy-Weinberg equilibrium. This assumption can be defended by the fact that a population that is possibly in disequilibrium reaches Hardy-Weinberg equilibrium in one round of random mating. We assume that there is no mutation.

2.1 Lemma. *A population of children formed by random mating from an arbitrary population of parents is in Hardy-Weinberg equilibrium at every autosomal locus, with allele relative frequencies equal to the allele relative frequencies in the population of the alleles of all parents.*

Proof. Let $p_{i,j}$ be the relative frequency of the ordered genotype (A_i, A_j) in the parents' population. Under random mating we choose a random father and independently a random mother, and each parent segregates a random allele to the child either his/her paternal or his/her maternal one. Given that the father segregates his paternal allele, he segregates A_i if and only if the father has genotype (A_i, A_j) for some j, which has probability $p_{i\cdot} = \sum_j p_{i,j}$. Given that the father segregates his maternal allele, he segregates A_i with probability $p_{\cdot i} = \sum_j p_{j,i}$. Therefore, the paternal allele of the child is A_i with probability

$$p_i' := \tfrac{1}{2}p_{i\cdot} + \tfrac{1}{2}p_{\cdot i}.$$

The mother acts in the same way and independently from the father. It follows that the child possesses ordered genotype (A_i, A_j) with probability $p_i'p_j'$.

Hence the children's population is in Hardy-Weinberg equilibrium. The probability p_i' is indeed the allele relative frequency of the allele A_i in the population of all parents. ∎

Hardy-Weinberg equilibrium is truly an *equilibrium*, in the sense that it is retained by further rounds of random mating. This follows from the lemma, because random mating produces Hardy-Weinberg equilibrium (so keeps it if it already present) and keeps the allele relative frequencies the same.

If the assumption of random mating is not satisfied, then Hardy-Weinberg equilibrium can easily fail. Population structure can lead to stable populations that are not in equilibrium, while selection may lead to fluctuations in allele frequencies. Random drift is another possible reason for deviations of Hardy-Weinberg equilibrium. In a particularly bad case of random drift an allele may even disappear from one generation to another, because it is not segregated by any parent, and of course can never come back.

2.2.1 Testing Hardy-Weinberg Equilibrium

To test Hardy-Weinberg equilibrium at a marker location, with alleles $A_1, \ldots, A_k$, we might take a random sample of n individuals and determine for each genotype A_iA_j the number N_{ij} of individuals in the sample with this genotype. We wish to test the null hypothesis that the probabilities of these genotypes factorize in the marginal frequencies of the alleles.

If parental and maternal origins of the alleles can be ascertained, then we can understand these numbers as referring to ordered genotypes (A_i, A_j). The frequencies N_{ij} then form a $(k \times k)$-table, and the null hypothesis asserts independence in this table, exactly as discussed for a standard test of independence in Section 14.1.5. A difference is that the marginal probabilities for the two margins (the allele frequencies) are a-priori known to be equal and hence the table probabilities p_{ij} are symmetric under the null hypothesis.

In a more realistic scenario the counts N_{ij} are the numbers of unordered genotypes $\{A_i, A_j\}$, which we can restrict to $i \leq j$ and provide half of a $(k \times k)$-table. Hardy-Weinberg equilibrium is that this half-table $N = (N_{ij})$ is multinomially distributed with parameters n and p_{ij} satisfying the relations $p_{ii} = \alpha_i^2$ and $p_{ij} = 2\alpha_i\alpha_j$ for $i < j$ and a probability vector (α_i). Thus the full parameter space is the unit simplex in the $\frac{1}{2}k(k+1)$-dimensional space, and the null hypothesis is a $k-1$-dimensional surface in this space. The null hypothesis can be tested by the chisquare or likelihood ratio test on $\frac{1}{2}k(k+1) - 1 - (k-1)$ degrees of freedom. The maximum likelihood estimator under the null hypothesis is the vector $(\hat\alpha_1, \ldots, \hat\alpha_k)$ of relative frequencies of the alleles $A_1, \ldots, A_k$ among the $2n$ measured alleles.

* 2.2.2 Estimating Allele Frequencies

Consider estimating the allele frequencies in a population, for a causal gene that is assumed to be the sole determinant of some phenotype. The data are a random sample from the population, and we assume Hardy-Weinberg equilibrium.

For codominant alleles this is easy. By the definition of codominance the two alleles of each individual can be determined from their (observed) phenotype and hence we observe the total numbers $N_1, \ldots, N_k$ of alleles $A_1, \ldots, A_k$. Under random sampling (with replacement) the distribution of the vector $(N_1, \ldots, N_k)$ is multinomial with parameters $2n$ and $p = (p_1, \ldots, p_k)$ and hence the maximum likelihood estimator of p_i is $N_i/(2n)$. In particular, the situation of codominance pertains for marker alleles, which are themselves observed.

On the other hand, if some alleles are recessive, then the numbers of alleles of the various types cannot be unambigously determined from the phenotypes, and hence the empirical estimators $N_i/(2n)$ are unavailable. Instead we observe for each possible phenotype the total number X_s of individuals with this phenotype; we assume that there are finitely many phenotypes, say $s = 1, \ldots, l$. Each phenotype is caused by a set of ordered genotypes (A_i, A_j) and hence the observational vector $(X_1, \ldots, X_l)$ is multinomially distributed with parameters n and $q = (q_1, \ldots, q_l)$, where q_s is the sum of the probabilities of the ordered genotypes that lead to phenotype s. Under Hardy-Weinberg equilibrium the probability of (A_i, A_j) is p_ip_j

and hence $q_s = \sum_{(i,j)\in s} p_i p_j$, where "$(i,j) \in s$" means that the ordered genotype (A_i, A_j) causes phenotype s. The likelihood for the observed data is therefore

$$(p_1, \ldots, p_k) \mapsto \binom{n}{X_1, \ldots, X_l} \prod_{s=1}^{l} \Bigl(\sum_{(i,j)\in s} p_i p_j \Bigr)^{X_s}.$$

We may maximize this over $p = (p_1, \ldots, p_k)$ to find the maximum likelihood estimators of the allele frequencies. If the grouping of the genotypes is not known, we may also maximize over the various groupings.

The maximization may be performed by the EM-algorithm, where we can choose the "full data" equal to the numbers $Y = (Y_{i,j})$ of individuals with ordered genotype (A_i, A_j). (Dropping the ordering is possible too, but do not forget to put factors 2 (only) in the appropriate places.) The full likelihood is, with $Y = (Y_{i,j})$,

$$(p_1, \ldots, p_k) \mapsto \binom{n}{Y} \prod_{(i,j)} (p_i p_j)^{Y_{i,j}}.$$

The observed data is obtained from the full data through the relationship $X_s = \sum_{(i,j)\in s} Y_{i,j}$. The EM-algorithm recursively computes

$$\begin{aligned} p^{(r+1)} &= \operatorname*{argmax}_p \mathrm{E}_{p^{(r)}} \Bigl(\sum_{(i,j)} Y_{i,j} \log(p_i p_j) |\, X_1, \ldots, X_l \Bigr) \\ &= \operatorname*{argmax}_p \sum_{(i,j)} X_{s(i,j)} \frac{p_i^{(r)} p_j^{(r)}}{\sum_{(i',j')\in s(i,j)} p_{i'}^{(r)} p_{j'}^{(r)}} \log(p_i p_j). \end{aligned}$$

Here $s(i,j)$ is the group s to which (i,j) belongs. The second equality follows because the conditional distribution of a multinomial vector Y given a set of totals $\sum_{h\in s} Y_h$ of subgroups is equal to the distribution of a set of independent multinomial vectors, one for each subgroup s, with parameters the total number of individuals in the subgroup and probability vector proportional to the original probabilities, renormalized to a probability vector for each subgroup. See the lemma below.

The target function in the final argmax is a linear combination of the values $\log p_i + \log p_j$, and is easier to maximize than the original likelihood in which the $p_i p_j$ enter through sums. Indeed, the argmax can be determined analytically. Under the assumption that the ordered genotypes $A_i A_j$ and $A_j A_i$ produce the same phenotype, i.e. $s(i,j) = s(j,i)$, we find, for $m = 1, \ldots, k$,

$$p_m^{(r+1)} \propto \sum_j X_{s(m,j)} \frac{p_m^{(r)} p_j^{(r)}}{\sum_{(i',j')\in s(m,j)} p_{i'}^{(r)} p_{j'}^{(r)}}.$$

The EM-algorithm iterates the corresponding equations to convergence.

2.2 Lemma. *Let $N=(N_1,\ldots,N_k)$ be multinomially distributed with parameter $(n,p_1,\ldots,p_k)$. Let $\{1,\ldots,k\}=\cup_{j=1}^l I_j$ be an arbitrary partition in l sets and let $M_j=\sum_{i\in I_j} N_i$ for $j=1,\ldots,l$. Then the conditional distribution of N given M is equal to the distribution of $N'=(N_1',\ldots,N_k')$ for*

(i) $(N_i' : i\in I_1),\ldots,(N_i' : i\in I_l)$ are independent.

(ii) $(N_i' : i\in I_j)$ is multinomially distributed with parameters (M_j,p_j') for $(p_j')_i = p_i/\sum_{i\in I_j} p_i$.

2.3 EXERCISE. Verify the claims in the preceding example and work out an explicit formula for the recursions.

* 2.2.3 Sex-linked Loci

The evolution of genotype frequencies for loci on the sex-chromosomes differs from that on the autosomes. Under random mating Hardy-Weinberg equilibrium is approached rapidly, but is not necessarily reached in finitely many matings. Of course, we need to make a difference between a male and a female population.

Consider a locus on the X-chromosome with possible alleles $A_1,\ldots,A_k$, and let p_i be the relative frequency of the allele A_i on the X-chromosome of the population of males, and let $Q_{i,j}$ be the relative frequency of the genotype (A_i,A_j), ordered by paternal origin (father, mother), on the two X-chromosomes in the population of females. Then $q_i=\frac12(\sum_j Q_{i,j}+\sum_j Q_{j,i})$ is the relative frequency of allele A_i in the population of females.

2.4 Lemma. *The relative frequencies in the population of children formed by random mating satisfy*

$$p_i'=q_i,\qquad Q_{i,j}'=p_iq_j,\qquad q_i'=\tfrac12(p_i+q_i).$$

Proof. A male descendant receive his X-chromosome from his mother, who segregates a random choice of her pair of alleles. If the mother is chosen at random from the population, then the allele is a random choice from the female alleles. This gives the first equation. For a female descendant to have genotype (A_i,A_j) her father must segregate A_i and her mother A_j. Under random mating this gives a choice of a random allele from the males and an independent choice of a random allele from the females. This proves the second assertion. The third assertion proved by computing q_i' from $Q_{i,j}'$. ■

Under the random mating assumption the alleles of a randomly chosen female are independent, but the relative frequencies $Q_{i,j}'$ are not symmetric in (i,j) as long as the male and female allele frequencies are different. This deviates from Hardy-Weinberg equilibrium. The formulas show that

$$p_i'-q_i'=-\tfrac12(p_i-q_i).$$

Thus if the male and female allele frequencies in the initial population are different, then they are different in all successive populations, and the difference alternates in sign. The difference converges exponentially fast to zero, and hence the female population rapidly approaches Hardy-Weinberg equilibrium.

For practical purposes under random mating the female population can be assumed to be in Hardy-Weinberg equilibrium. One consequence is that the prevalence of diseases that are caused by a single recessive gene on the X-chromosome is much higher in males than in females (under the assumption that the disease will appear as soon as a male has the causal variant on his X-chromosome).

2.3 Linkage Equilibrium

Whereas Hardy-Weinberg equilibrium refers to the two alleles at a single locus, linkage equilibrium refers to the combination of multiple loci in a single haplotype. A population is said to be in *linkage equilibrium* (LE) if the k alleles on a k-loci haplotype that is chosen at random from the population are independent.

For a more precise description in the case of two-loci haplotypes, suppose that the two loci have k and l possible alleles, $A_1, \ldots, A_k$ and $B_1, \ldots, B_l$, respectively. Then there are kl possible haplotypes for the two loci: every combination A_iB_j for $i = 1, \ldots, k$ and $j = 1, \ldots, l$. A population is said to be in *linkage equilibrium* at the two loci if a randomly chosen haplotype from the population is A_iB_j with probability p_iq_j, where p_i and q_j are the probabilities that a randomly chosen allele at the first or second locus is A_i or B_j, respectively. Here the “population of haplotypes” should be understood as the set of all haplotypes of individuals, each individual contributing two haplotypes and the haplotypes being stripped from information on their origin.

Unlike Hardy-Weinberg equilibrium, linkage equilibrium does not necessarily arise after a single round of (random) mating. The reason is that in the segregation process whole pieces of chromosome are passed on, rather than individual loci. Because crossovers, which delimit the pieces of chromosome that are passed on intact, occur on the average only 1–2 times per chromosome and at more or less random loci, it is clear that between loci that are close together linkage equilibrium can at best be reached after many generations. This can be made precise in terms of the recombination fraction between the loci.

Consider the following schematic model for the formation of two-locus gametes. We draw two haplotypes at random from an existing population of haplotypes and next form an offspring of one haplotype out of these in two steps:

(i)-1 passing the original pair of haplotypes on unchanged with probability $1 - \theta$,
(i)-2 cutting and recombining the haplotypes with probability θ.
(ii) picking one of the two resulting haplotypes at random.

To form a new population of two-locus haplotypes, we repeat this experiment infinitely often, independently.

Let h_{ij} be the relative frequency of haplotype A_iB_j in the initial population, and let $p_i = \sum_j h_{ij}$ and $q_j = \sum_i h_{ij}$ the corresponding relative frequencies of the alleles A_i and B_j, respectively.

2.5 Lemma. *The relative frequency h'_{ij} of haplotype A_iB_j in the new population produced by scheme (i)-(ii) satisfies*

$$h'_{ij} - p_iq_j = (1-\theta)(h_{ij} - p_iq_j).$$

The corresponding marginal relative frequencies of the alleles A_i and B_j are $p'_i = p_i$ and $q'_j = q_j$.

Proof. Consider a haplotype formed by the chance mechanism as described in (i)-(ii). If R is the event that recombination occurs, as in (i)-2, then the probability that the haplotype is A_iB_j is

$$h'_{ij} = P(A_iB_j\,|\,R^c)P(R^c) + P(A_iB_j\,|\,R)P(R) = h_{ij}(1-\theta) + p_iq_j\theta.$$

Here the second equality follows, because in the absence of recombination, as in (i)-1, the haplotypes that are passed on are identical to the originally chosen haplotypes, while given recombination the haplotype A_iB_j is passed on if it is reconstituted from a pair of original haplotypes of the forms A_iB_s and A_tB_j for some s and t, which have frequencies p_i and q_j, respectively.

By summing the preceding display over i or j, we find that marginal relative frequencies of the alleles in the new population are equal to the marginal relative frequencies p_i and q_j in the initial population. Next the lemma follows by rearranging the preceding display. ■

By repeated application of the lemma we see that the haplotype relative frequencies $h_{ij}^{(n)}$ after n rounds of mating satisfy

$$h_{ij}^{(n)} - p_iq_j = (1-\theta)^n(h_{ij} - p_iq_j).$$

This implies that $h_{ij}^{(n)} \to p_iq_j$ as $n \to \infty$ provided that $\theta > 0$, meaning that the population approaches linkage equilibrium. The speed of convergence is exponential for any $\theta > 0$. However, if the two loci are tightly linked, then $1-\theta \approx 1$ and the convergence is slow. If the two loci are unlinked, then $1-\theta = \frac{1}{2}$ and the population approaches equilibrium quickly. The convergence also depends on the *linkage disequilibrium* parameter $D_{ij} = h_{ij} - p_iq_j$ in the initial population. The lemma says precisely that the disequilibrium in the new population satisfies $D'_{ij} = (1-\theta)D_{ij}$.

The preceding scheme (i)-(ii) produces only one haplotype, while individuals consists of two haplotypes. We may view of (i)-(ii) as the model for choosing one parent from the population and the single gamete (s)he produces by a meiosis. Under random mating the parents are chosen independently from the population and, as usual, we assume all meioses independent. Therefore, under random mating the scheme can be lifted to the level of diploid individuals by creating the new population as pairs of independent haplotypes, both produced according to (i)-(ii). Actually, the start of scheme (i)-(ii) by choosing two *independent* haplotypes already reflects the implicit assumption that the haplotypes of the individuals in the parents' population are independent. This is not an important loss of generality, because independence will arise after one round of random mating.

The lemma extends without surprises to multi-loci haplotypes. This is discussed in the more general set-up that includes selection in Section 2.6.

2.4 Full Equilibrium

Linkage equilibrium as defined in Section 2.3 refers to haplotypes, and does not prescribe the distribution of genotypes, which are pairs of haplotypes. We define a population to be in *combined Hardy-Weinberg and linkage equilibrium*, or simply in *equilibrium*, if the population is in both Hardy-Weinberg and linkage equilibrium *and* the two haplotypes within the genotype of a randomly chosen individual are independent. Thus the combination of HW and LE is more than the union of its constituents (except in the case of a single locus, where LE is empty).

We shall also refer to independence of the two haplotypes of an arbitrary individual as *Hardy-Weinberg at the haplotype level.* This is ensured by random mating: knowing one haplotype of a person gives information about one parent, but under random mating is not informative about the other parent, and hence the other haplotype. This is perhaps too obvious to say, and that is why the assumption is often not explicitly stated. Hardy-Weinberg at the haplotype level is in general not an equilibrium.

Note the two very different causes of randomness and (in)dependence involved in Hardy-Weinberg equilibrium and linkage equilibrium: Hardy-Weinberg equilibrium results from mating, which is at the population level, whereas linkage equilibrium results from meiosis, which is at the cell level. Furthermore, even if under random mating a child can be viewed as the combination of two independent haplotypes, these haplotypes are formed possibly by recombination, and under just random mating cannot be viewed as physically drawn at random from some population of haplotypes.

A concrete description of combined Hardy-Weinberg and linkage equilibrium for two-loci haplotypes is as follows. Let $p_1, \ldots, p_k$ and $q_1, \ldots, q_l$ be the population fractions of the alleles $A_1, \ldots, A_k$ at the first locus and the alleles $B_1, \ldots, B_l$ at the second locus, respectively. If the population is in combined Hardy-Weinberg and linkage equilibrium, then an ordered genotype of an arbitrary person consists of haplotypes A_iB_j and $A_{i'}B_{j'}$ with probability $p_ip_{i'}q_jq_{j'}$. Thus a genotype is formed by independently constructing two haplotypes by glueing together alleles that are chosen independently according to their population frequencies.

Here we considered ordered genotypes $(A_iB_j, A_{i'}B_{j'})$. Stripping the order would introduce factors 2.

* 2.5 Population Structure

Population structure is a frequent cause for deviations from equilibrium. Consider for instance a population consisting of several subpopulations, each of which satisfies the random mating assumption within itself, but where there are no interactions between the subpopulations. After one round of random mating each subpopulation will be in Hardy-Weinberg equilibrium. However, unless the allele frequencies for the subpopulations are the same, the population as a whole will not be in Hardy-Weinberg equilibrium. Similarly each of the subpopulations, but not the whole population, may be in equilibrium.

This is shown in the following lemma, which applies both to single loci and haplotypes. It is assumed that there are k different haplotypes $A_1, \ldots, A_k$, and within each subpopulation the individuals consist of random combinations of two haplotypes. The lemma shows that individuals in the full population are typically not random combinations of haplotypes.

The lemma is a special case of the phenomenon that two variables can be conditionally uncorrelated (or independent) given a third variable, but unconditionally correlated (or dependent). The proof of the lemma is based on the general rule, valid for any three random variables X, Y, N defined on a single probability space,

$$\text{(2.6)} \qquad \operatorname{cov}(X,Y) = \mathrm{E}\operatorname{cov}(X,Y|\,N) + \operatorname{cov}\bigl(\mathrm{E}(X|\,N), \mathrm{E}(Y|\,N)\bigr).$$

2.7 Lemma. *Consider a population consisting of N subpopulations, of fractions $\lambda_1, \ldots, \lambda_N$ of the full population, each of which is in Hardy-Weinberg at the haplotype level, the nth subpopulation being characterized by the relative frequencies $p_1^n, \ldots, p_k^n$ of the haplotypes $A_1, \ldots, A_k$. Then the relative frequency $p_{i,j}$ of the ordered genotype (A_i, A_j) in the full population satisfies $\sum_j p_{i,j} = \sum_j p_{j,i} =: p_{i\cdot}$, and*

$$p_{i,j} - p_{i\cdot}p_{j\cdot} = \sum_{n=1}^N \lambda_n (p_i^n - p_{i\cdot})(p_j^n - p_{j\cdot}).$$

Proof. We apply (2.6) with the variable X equal to 1 or 0 if the paternal haplotype of a randomly chosen individual from the full population is A_i or not, with the variable Y defined similarly relative to the maternal haplotype and j instead of i, and with N the index of the subpopulation that the individual belongs to. Then $\mathrm{E}X = p_{i\cdot}$ is the relative frequency of haplotype A_i in the full population, $\mathrm{E}(X|\,N = n) = p_i^n$, the variable Y satisfies the same equalities, but with j instead of i, and $P(N = n) = \lambda_n$ for every n. The mean $\mathrm{E}X = \mathrm{E}\mathrm{E}(X|\,N) = \sum_n \lambda_n p_i^n$ is equal to the relative frequency of the paternal allele A_i in the population, and similarly $\mathrm{E}Y = \sum_n \lambda_n p_j^n$ is the relative frequency of a maternal allele A_j in the population.

These relative frequencies can also be written $\sum_k p_{k,i}$ and $\sum_k p_{k,j}$, respectively. Choosing $i = j$, we have $\mathrm{E}X = \mathrm{E}Y$ and hence the paternal and maternal allele relative frequencies coincide, showing that $\sum_k p_{k,i} = \sum_k p_{k,j}$. To prove the validity of the display the second we note that $\mathrm{cov}(X, Y)$ is the left side of the lemma and apply (2.6). The assumption of independence of haplotypes in each subpopulation shows that $\mathrm{cov}(X, Y|\, N) = 0$, so that the first term on the right side of the preceding display vanishes. The second term is the right side of the lemma. ■

Taking $i = j$ in the lemma, we obtain that the relative frequency $p_{i,i}$ of the homozygous individuals (A_i, A_i) satisfies

$$p_{i,i} - p_{i\cdot}^2 = \sum_{n=1}^{N} \lambda_n (p_i^n - p_{i\cdot})^2 \geq 0.$$

The expression is strictly positive unless the relative frequency of A_i is the same in every subpopulation. The inequality shows that the proportion of homozygous individuals is larger than it would be under Hardy-Weinberg equilibrium in the full population: the *heterozygosity* $1 - \sum_i p_{i,i}$ is smaller than its value $1 - \sum_i p_{i\cdot}^2$ under Hardy-Weinberg.

* 2.6 Viability Selection

A population is under *selection* if not every individual or every mating pair has the same chance to produce offspring. Selection changes the composition of future generations. The genotypes in successive generations may still tend to an equilibrium, but they may also fluctuate forever.

The simplest form of selection is *viability selection.* This takes place at the level of individuals and can be thought of as changing an individual's chances to "survive" until mating time and produce offspring. Viability selection is modelled by attaching to each genotype (A_i, A_j) a measure $w_{i,j}$ of *fitness.* Rather than choosing a parent at random from the population (according to the population relative frequency $p_{i,j}$ for genotype (A_i, A_j)), we choose the parent (A_i, A_j) with probability proportional to $p_{i,j} w_{i,j}$. For simplicity we assume that $w_{i,j}$ is symmetric in (i, j), nonnegative and not identically zero.

We retain a random mating assumption in that we independently choose two parents according to this mechanism. Each of the parents segregates one gamete by a meiosis, and these combine into a zygote. We assume that there are no mutations. The children's population will consist of pairs of independently produced haplotypes, and hence be in "Hardy-Weinberg equilibrium at the haplotype level". In this situation it suffices to study the haplotype frequencies of the gametes produced in a single meisosis. It is also not a serious loss of generality to assume that the relative frequency of genotype (A_i, A_j) in the initial parents' population factorizes as $p_i p_j$, for p_i the marginal relative frequency of haplotype p_i. The relative frequency of a child-bearing parent (A_i, A_j) is then proportional to $p_i p_j w_{i,j}$. The proportionality factor is the inverse of the sum over all these numbers

$$F(p) = \sum_i \sum_j w_{i,j} p_i p_j.$$

The value $F(p)$ is the *average fitness* of a population characterized by the allele relative frequencies p. For a given haplotype i we also define the *marginal fitness* of allele A_i by

$$F_i(p) = \sum_j p_j w_{i,j}.$$

This can also be viewed as the expected fitness of an individual who is known to possess at least one allele A_i.

For single-locus genotypes and a fitness measure that remains the same over time, the evolution of the population under viability selection can be easily summarized: the fitness of the populations increases and the relative frequencies typically tend to an equilibrium. On the other hand, for multiple loci haplotypes, or fitness that depends on the composition of the population, the situation is already complicated and many types of behaviour are possible, including cyclic behaviour.

2.6.1 Single Locus

For a single locus genotype meiosis just consists of the segregating parent choosing one allele from his pair of alleles at random. We assume that the parents' population is in Hardy-Weinberg equilibrium and write $p_1, \ldots, p_k$ for the marginal relative frequencies of the possible alleles $A_1, \ldots, A_k$. A segregating father has genotype (A_i, A_j) with probability proportional to $p_i p_j w_{i,j}$, and hence the paternal allele of a child is A_i with probability p_i' satisfying

$$p_i' \propto \tfrac{1}{2} \sum_j p_i p_j w_{i,j} + \tfrac{1}{2} \sum_j p_j p_i w_{j,i} = p_i F_i(p). \tag{2.8}$$

The proportionality factor is the fitness $F(p)$ of the population. Equations (2.8) show immediately that a frequency vector $p = (p_1, \ldots p_k)^T$ is a fixed point of the iterations ($p' = p$) if and only if the marginal fitnesses of all alleles with $p_i > 0$ are the same. The equation also shows that once an allele has disappeared, then it will never come back ($p_i' = 0$ whenever $p_i = 0$, for any i).

By straightforward algebra (see the proof below) it can be derived that

$$p' - p = \frac{1}{2F(p)} \bigl(\operatorname{diag}(p) - pp^T\bigr) \nabla F(p). \tag{2.9}$$

Because the gradient $\nabla F(p)$ is the direction of maximal increase of the fitness function F, this equation suggests that the iterations (2.8) "attempt to increase the fitness". In the next theorem it is shown that, indeed, successive populations become ever fitter; as geneticists phrase it: "the populations climb the fitness surface".

The presence of the matrix $\mathrm{diag}\,(p) - pp^T$ makes the preceding display somewhat difficult to interpret. If all coordinates of p are positive, then the null space of the matrix is the linear span of the constant vector 1. Hence fixed points of the iteration ($p' - p = 0$) in the interior of the unit simplex are characterized by $\nabla F(p) \propto 1$, which is precisely the Lagrange equation for an extremum of the function F under the side condition $p^T 1 = 1$ that p is a probability vector. However, minima and saddlepoints of the Lagrangian will also set the right side of (2.9) to zero, so that the equation suggests maximization of fitness, but does not prove it.

We show in the next theorem that the sequence of iterates $p, p', p'', \ldots$ (2.8) converges from any starting vector p to a limit. The limit is necessarily a fixed point, which may have some coordinates equal to 0. In the following theorem we concentrate on the most interesting case that the limit is an interior point of the unit simplex, so that all alleles are present.

2.10 Theorem. *The iterations (2.8) can be written in the form (2.9) and satisfy $F(p') \geq F(p)$, for any p, with equality only if $p' = p$. Any sequence of iterates $p, p', p'', \ldots$ converges. If the limit is in the interior of the unit simplex, then the convergence is exponentially fast.*

2.11 Theorem. *The interior $\mathring{S}$ of the unit simplex is attracted by a single vector in the $\mathring{S}$ if and only if F assumes its global maximum uniquely at a point in $\mathring{S}$, where the point of maximum is the attractor. A necessary and sufficient condition for this to happen is that the matrix $(w_{i,j})$ has one strictly positive and $k-1$ strictly negative eigenvalues and that there exists a fixed point in $\mathring{S}$.*

Proofs. For W the matrix $(w_{i,j})$, the fitness function is the quadratic form $F(p) = p^T W p$. The recursion (2.8) for the allele relative frequencies can be written in the matrix form $p' = \mathrm{diag}\,(p) W p / F(p)$, and hence

$$p' - p = \frac{\mathrm{diag}\,(p) W p - p\,F(p)}{F(p)} = \frac{\mathrm{diag}\,(p) W p - pp^T W p}{F(p)}.$$

As $\nabla F(p) = 2Wp$, the right side is the same as in (2.9).

Inserting the recursion (2.8) into $F(p') = \sum_i \sum_j p_i' p_j' w_{i,j}$, we see that

$$F(p)^2\, F(p') = \sum_i \sum_j p_i p_j F_i(p) F_j(p) w_{i,j} = \sum_i \sum_j \sum_k p_i p_j p_k F_i(p) w_{i,j} w_{j,k}.$$

Because the product $p_i p_j p_k w_{i,j} w_{j,k}$ is symmetric in (i,k), we can replace $F_i(p)$ in the right side by $\bigl(F_i(p) + F_k(p)\bigr)/2$, which is bigger than $\sqrt{F_i(p)}\sqrt{F_k(p)}$. Thus the preceding display is bigger than

$$\begin{aligned}\sum_i \sum_j \sum_k p_i p_j p_k \sqrt{F_i(p)}\sqrt{F_k(p)} w_{i,j} w_{j,k} &= \sum_j p_j \Bigl(\sum_i p_i \sqrt{F_i(p)} w_{i,j}\Bigr)^2 \\ \geq \Bigl(\sum_j \sum_i p_j p_i \sqrt{F_i(p)} w_{i,j}\Bigr)^2 &= \Bigl(\sum_i p_i F_i(p)^{3/2}\Bigr)^2.\end{aligned}$$

By Jensen's inequality applied to the convex function $x \mapsto x^{3/2}$, this is bigger than $\bigl(\sum_i p_i F_i(p)\bigr)^3 = F(p)^3$. We divide by $F(p)^2$ to conclude the proof that the fitness is nondecreasing.

By application of Jensen's inequality with a second order term, the last step of the preceding derivation can be refined to yield the inequality $\sum_i p_i F_i(p)^{3/2} \geq F(p)^{3/2} + C\sigma^2(p)$, for $\sigma^2(p) = \sum_i p_i\bigl(F_i(p) - F(p)\bigr)^2$ the variance of the marginal fitness, and C the minimum of the second derivative of $x \mapsto x^{3/2}$ on the convex hull of the marginal fitnesses ($C = (3/8)(\max_{i,j} w_{i.j})^{-1/2}$ will do). Insertion of this improved bound, we see that, for any $p \in S$,

$$F(p') - F(p) \gtrsim \sigma^2(p), \tag{2.12}$$

In particular, $F(p') > F(p)$ unless all marginal frequencies are equal, in which case $p' = p$ by (2.8).

By the compactness of the unit simplex, any sequence of iterates p^n of (2.8) possesses limit points. Because the sequence $F(p^n)$ is increasing, the fitness $F(p^*)$ of all limit points is the same. If $p^{n_j} \to p^*$, then $p^{n_j+1} \to (p^*)'$, and hence $(p^*)'$ is a limit point as well. Therefore, $F(p^*) = F(p^*)')$, whence $p^* = (p^*)'$, showing that any limit point is necessarily a fixed point of the iteration.

To prove that each sequence actually has only a single limit point, we derive below that, for any fixed point p^* of the iterations and any p sufficiently close to p^*,

$$F(p^*) - F(p) \lesssim \sigma^{4/3}(p). \tag{2.13}$$

By (2.8) and the Cauchy-Schwarz inequality,

$$\|p^{n+1} - p^n\|_1 = \frac{1}{F(p^n)} \sum_i p_i \bigl|F_i(p^n) - F(p^n)\bigr| \lesssim \sigma(p^n).$$

We rewrite the right side as $\sigma^2(p^n)/\sigma(p^n)$ and (2.12) and (2.13) to see that the right side is bounded above by, if p^n is sufficiently close to a fixed point p^*,

$$\frac{F(p^{n+1}) - F(p^n)}{\bigl(F(p^*) - F(p^n)\bigr)^{3/4}} \lesssim \bigl(F(p^*) - F(p^n)\bigr)^{1/4} - \bigl(F(p^*) - F(p^{n+1})\bigr)^{1/4}.$$

If $p^{n+1}, p^{n+2}, \ldots$ are again in the neighbourhood of p^* where (2.13) is valid, then we can repeat the argument and find that

$$\|p^{n+k} - p^n\|_1 \leq \sum_{i=1}^k \|p^{n+i} - p^{n+i-1}\|_1 \lesssim \bigl(F(p^*) - F(p^n)\bigr)^{1/4}, \tag{2.14}$$

by telescoping of the sum of upper bounds. Now if (2.13) is valid for $\|p - p^*\| < \varepsilon$ and we start with p^n such that both $\|p^n - p^*\|_1$ and the right side of the preceding display are smaller than $\delta \leq \varepsilon/2$, then the sequence $p^{n+1}, p^{n+2}, \ldots$ will remain within distance ε of p^* and hence will satisfy the preceding display, which then shows that $\|p^{n+k} - p^n\|_1 < \delta$ for all k. Because δ is arbitrary this shows that the sequence is Cauchy and hence has a single limit point.

A point $p^* \in \mathring{S}$ is a fixed point if and only if $F_i(p^*) = F(p^*)$ for every i, or equivalently $\nabla F(p^*) \propto 1$. This implies that $F(p) - F(p^*) = F(p - p^*)$ and $\sigma^2(p) = \sum_i p_i F_i^2(p - p^*) - F^2(p - p^*)$, for every p. Also, for p sufficiently close to p^* the coordinates of p are bounded away from zero, and hence

$$\sigma^2(p) \gtrsim \sum_i F_i^2(p - p^*) - F^2(p - p^*) \gtrsim F(p - p^*) - F^2(p - p^*),$$

because $\|Wv\|^2 \geq Cv^T Wv$ for any symmetric matrix W and v, and C the smallest strictly positive eigenvalue of W. This proves that $F(p^*) - F(p) \lesssim \sigma^2(p)$, for p sufficiently close to p^*. We combine this with (2.12) to see that if p^n tends to $p^* \in \mathring{S}$, then

$$F(p^*) - F(p^n) \lesssim \sigma^2(p^n) \lesssim F(p^{n+1}) - F(p^n).$$

This inequality can be rearranged to see that $F(p^*) - F(p^{n+1}) \leq C\bigl(F(p^*) - F(p^n)\bigr)$, for some $C < 1$. Consequently, the sequence $F(p^*) - F(p^n)$ tends to zero exponentially fast, and hence so does the sequence $\|p^* - p^n\|$, by (2.14).

The proof of (2.13) is based on a similar argument applied to the projection $\bar{p}$ of a given p on the set $S_I := \{p \in S : p_i = 0 \forall i \notin I\}$, where $I = \{i : F_i(p^*) = F(p^*)\}$. Because $p^* \in S_I$ we now have $F(\bar{p}) - F(p^*) = F(\bar{p} - p^*)$ and $\sigma^2(\bar{p}) = \sum_i \bar{p}_i F_i^2(\bar{p} - p^*) - F^2(\bar{p} - p^*)$, and, for p close to p^*,

$$\sigma^2(\bar{p}) \gtrsim \sum_{i \in I} |\bar{p}_i - p_i^*| F_i^2(\bar{p} - p^*) - F^2(\bar{p} - p^*) \gtrsim |F(\bar{p} - p^*)|^{3/2} - F^2(\bar{p} - p^*),$$

because $\sum_i |v_i|(Wv)_i^2 \gtrsim |v^T Wv|^{3/2}$ for any symmetric matrix W.[♯] We also have that,

$$\sigma^2(p) - \sigma^2(\bar{p}) = \nabla\sigma^2(\tilde{p})(p - \bar{p}) \geq c \sum_{i \in I} p_i - C \sum_{i \notin I} |p_i - \bar{p}_i|,$$

for c and C the minimum and maximal value of $\partial\sigma^2/\partial p_i$ over the convex segment between $\bar{p}$ and p^* and $i \notin I$ and $i \in I$, respectively. Because $\partial\sigma^2/\partial p_i(p^*) = \bigl(F_i(p^*) - F(p^*)\bigr)^2$, for p sufficiently close to p^* we can choose c bounded away from 0 and C arbitrarily close to 0. It follows that

$$\sigma^2(p) - \sigma^2(\bar{p}) \gtrsim \sum_{i \notin I} p_i - \varepsilon \sum_{i \in I} |p_i - \bar{p}_i| \gtrsim \|p - \bar{p}\|_1 \gtrsim \bigl|F(p) - F(\bar{p})\bigr|.$$

For the last inequality we use that $\bar{p}_i = p_i + |I|^{-1} \sum_{i \notin I} p_i$ for $i \in I$, and $\|p - \bar{p}\|_1 = 2\sum_{i \notin I} p_i$. The last display shows that $\sigma^2(\bar{p}) \leq \sigma^2(p)$. Finally

$$F(p^*) - F(p) = F(p^* - \bar{p}) + F(\bar{p}) - F(p) \lesssim \sigma^{4/3}(p) + \sigma^2(p).$$

This concludes the proof of (2.13), and hence of the first theorem.

A point P of global maximum of F is necessarily a fixed point, as otherwise $F(P') > F(P)$. A fixed point P in $\mathring{S}$ satisfies $WP \propto 1$ and then every vector v with

[♯] Lyubich,?? This more involved inequality is used, because not necessarily $p_i^* > 0$ for $i \in I$.

$Wv = 0$ satisfies $v^T 1 \propto v^T WP = 0$, by the symmetry of W. Thus the vector $P + \varepsilon v$ is contained in $\mathring{S}$ for sufficiently small $\varepsilon > 0$ and satisfies $W(P+\varepsilon v) = WP \propto 1$ and $(P+\varepsilon v)^T W(P+\varepsilon v) = P^T WP$. It follows P is a fixed point or is a unique point of maximum only if the kernel of W is trivial. We assume this in the following.

If p^* is a fixed point that is the limit of a sequence iterates p^n that starts in the interior, then $F_i(p^*) \leq F(p^*)$ for every i. Indeed $F(p^*) \geq F(p^1) > 0$ and hence $p_i^{n+1}/p_i^n = F_i(p^n)/F(p^n) \to F_i(p^*)/F(p^*)$. If the right side is bigger than 1, then eventually $p_i^{n+1} > cp_i^n$ for some $c > 1$, which is possible only if $p_i^n = 0$ eventually. This is impossible, as $p^n \in \mathring{S}$ for all n, as follows from the fact that $Wp > 0$ if $p > 0$ under the assumption that no row of W vanishes.

If $P \in \mathring{S}$ is a point of global maximum and p^* is the limit of a sequence of iterates starting in $\mathring{S}$, then $WP = F(P)1$ and hence

$$F(P) = \sum_i p_i^* F(P) = \sum_i p_i^* \sum_j w_{i,j} P_j = \sum_j P_j \sum_i w_{j,i} p_i^* = \sum_j P_j F_j(p^*) \leq F(p^*).$$

Hence p^* is also a point of global maximum. Therefore, if F has a unique point of global maximum, then every sequence of iterates that starts in the interior tends to P. Conversely, if every sequence of iterates tends to a point P in the interior, then P must be a unique point of global maximum as the iterations increase the value of F from any starting point.

Because W is symmetric, it is diagonalizable by an orthogonal transformation, with its eigenvalues as the diagonal elements. These eigenvalues are real and nonzero and hence can be written $\lambda_1 \geq \cdots \geq \lambda_l > 0 > \lambda_{l+1} \geq \cdots \geq \lambda_k$, for some $0 \leq l \leq k$. In fact $l \geq 1$, because otherwise W is negative-definite, contradicting that W has nonnegative elements and is nonzero. The function F has a unique point of maximum at a point P in $\mathring{S}$ if and only if $F(P+u) < F(P)$ for every u such that $P + u \geq 0$ and $u^T 1 = 0$. Because necessarily $WP \propto 1$, we have $F(P+u) = F(P) + u^T Wu$ for such u, and hence this is equivalent to $u^T Wu < 0$ for every $u \neq 0$ such that $P + u \geq 0$ and $u^T WP = 0$. If the diagonalizing transformation maps P to Q, u to v, and $\{u\colon P+u \geq 0\}$ to V, then this statement is equivalent to $\sum_i \lambda_i v_i^2 < 0$ for nonzero $v \in V$ such that $\sum_i \lambda_i v_i Q_i = 0$.

Because $P \in \mathring{S}$, the set V contains an open neighbourhood of 0. If $l \geq 2$, then there exist solutions $v \in V$ of the equation $\sum_i \lambda_i v_i Q_i = 0$ with $v_{l+1} = \cdots = v_k = 0$ and $(v_1, \ldots, v_l) \neq 0$, which is incompatible with the inequality $\sum_i \lambda_i v_i^2 < 0$. We conclude that $l < 2$ if F has a unique point of maximum in $\mathring{S}$.

Conversely, suppose that $l = 1$ and the iterations have a fixed point $P \in \mathring{S}$. The latter implies that $WP \propto 1$. If v solves the equation $\sum_i \lambda_i v_i Q_i = 0$, then by the Cauchy-Schwarz inequality,

$$|\lambda_1 v_1 Q_1|^2 = \Bigl|\sum_{i\geq 2} \lambda_i v_i Q_i\Bigr|^2 \leq \sum_{i\geq 2} |\lambda_i| v_i^2 \sum_{i\geq 2} |\lambda_i| Q_i^2.$$

Consequently, for any such v,

$$\sum_i \lambda_i v_i^2 \leq \Bigl(\frac{\sum_{i\geq 2} |\lambda_i| Q_i^2}{\lambda_1 Q_1^2} - 1\Bigr) \sum_{i\geq 2} |\lambda_i| v_i^2.$$

In terms of the old coordinates the left side is $u^T W u$. The first term on the right side is negative if $\sum_i \lambda_i Q_i^2 > 0$, which is true because this expression is equal to $P^T W P$. It follows that a fixed point $P \in \mathring{S}$ is automatically a unique point of global maximum of F. ■

2.15 Example (Two alleles). In the situation of two alleles ($k = 2$), the recursions can be expressed in the single probability $p_1 = 1 - p_2$. We have $F(p) = p_1 F_1(p) + p_2 F_2(p)$, for $F_i(p) = \sum_j w_{i,j} p_j$ the marginal fitness of allele A_i, and

$$p_1' - p_1 = p_1 p_2 \frac{F_1(p) - F_2(p)}{F(p)}.$$

Not surprisingly, the frequency of allele A_1 increases if the marginal fitness of A_1 in the current population exceeds the marginal fitness of A_2. The "fitness surface" F can be written as a quadratic function of p_1. The maximum fitness is in the interior of the interval $[0, 1]$ if the fitness parabola has its apex at a point in $(0, 1)$. In that case the population will tend to an equilibrium in which both alleles are present. The fitness parabola may also have one or two local maxima at the boundary points 0 and 1. In the latter case one of the two alleles will disappear, where it may depend on the starting point which one. □

2.16 EXERCISE. Consider a population of individuals (A_i, A_j) in Hardy-Weinberg with vector of allele relative frequencies $p = (p_1, \ldots, p_k)^T$. Define N_i to be the number of alleles A_i in a random person from this population, and let W be the fitness of this individual, so that $2(\text{diag}\,(p) - pp^T)$ is the covariance matrix of the (multinomial) vector $(N_1, \ldots, N_k)$, and $F(p) = \mathrm{E}_p W$. Show that $p' - p = (1/2\mathrm{E}_p W)\,\mathrm{cov}_p(W, N)$.

2.6.2 Multiple Loci

If the fitness depends on the genotype at multiple loci, then the recursions for the gamete frequencies incorporate recombination probabilities. Consider individuals (A_i, A_j) consisting of ordered pairs of two k-loci haplotypes A_i and A_j. We identify the haplotypes with sequences $i = i_1 i_2 \cdots i_k$ and $j = j_1 j_2 \cdots j_k$, where each i_s and j_s refers to a particular allele at locus s, and write $w_{i,j}$ for the fitness of individual (A_i, A_j), as before. For simplicity we assume that the fitness $w_{i,j}$ is the same for every pair (i, j) that gives the same unordered genotypes $\{i_1, j_1\}, \ldots, \{i_k, j_k\}$ at the k loci; in particular $w_{i,j} = w_{j,i}$. Thus the haplotype structure (i, j) is unimportant for the value of fitness, an assumption known as absence of *cis action*.

As before, we assume that the two haplotypes of a randomly chosen parent from the population are independent, and write p_i for the relative frequency of haplotype A_i. Thus a father of type (A_i, A_j) enters a mating pair and has offspring with probability proportional to $p_i p_j w_{i,j}$, where the proportionality constant is the fitness $F(p) = \sum_i \sum_j p_i p_j w_{i,j}$ of the parents' population. We also retain the notation $F_i(p) = \sum_j p_j w_{i,j}$ for the marginal fitness of haplotype i. We add the *lazy notation* p_{i_s} for the marginal relative frequency of allele i_s at locus s and, more generally, p_{i_S} for the relative frequency of the sub-haplotype defined by the alleles $i_S = (i_s\colon s \in S)$ inside $i = i_1 \cdots i_k$ at the loci $s \in S$, for $S \subset \{1, \ldots, k\}$. Thus, for $i = i_1 i_2 \cdots i_k$,

$$p_{i_S} = \sum_{j: j_S = i_S} p_j.$$

Thus the form of the subscript reveals the type of relative frequency involved.

A gamete produced by a parent of type (A_i, A_j) is of type A_h for $h = h_1 \cdots h_k$ a combination of indices chosen from $i = i_1 i_2 \cdots i_k$ and $j = j_1 j_2 \cdots j_k$, each h_s being equal to either i_s or j_s. With further abuse of notation we shall refer to this gamete by $i_S j_{S^c}$ if $S \subset \{1, 2 \ldots, k\}$ is the set of indices in h taken from i, and we write the corresponding haplotype frequency as $p_{i_S j_{S^c}}$. The set S in reality cuts the loci $1, \ldots, k$ into a number of groups of adjacent loci separated by elements of S^c, a fact that is poorly expressed in the notation $i_S j_{S^c}$. The probability c_S that a father of type (A_i, A_j) segregates a gamete of type $A_{i_S j_{S^c}}$ is equal to 1/2 times the probability of occurrence of recombination between every endpoint of a segment in S and starting point of a segment in S^c and vice versa. This probability can be derived from the model for the chiasmata process (see Theorem 1.3), but in this section we shall be content with the notation c_S.

A gamete produced by a given meiosis is of type $i = i_1 i_2 \cdots i_k$ if for some S the parent possesses and segregates paternal alleles i_S at loci S and maternal alleles i_{S^c} at loci S^c, or the other way around. To be able to do this the parent must be of type $(A_{i_S j_{S^c}}, A_{j_S i_{S^c}})$ for some j, or of type $(A_{j_S i_{S^c}}, A_{i_S j_{S^c}})$, respectively. This shows that the frequency p_i' of haplotype i in the children's population satisfies

$$p_i' \propto \sum_S c_S \Bigl(\tfrac{1}{2} \sum_j p_{i_S j_{S^c}} p_{j_S i_{S^c}} w_{i_S j_{S^c}, j_S i_{S^c}} + \tfrac{1}{2} \sum_j p_{j_S i_{S^c}} p_{i_S j_{S^c}} w_{j_S i_{S^c}, i_S j_{S^c}} \Bigr).$$

By the symmetry assumptions on the fitness matrix $(w_{i,j})$, the two sums within the brackets are equal (the fitness is $w_{,j}$ in the terms of both sums). Let $F_i(p) = \sum_j p_j w_{i,j}$ be the marginal fitness of haplotype A_i, and

$$D_{i,j}^S(p) = p_i p_j - p_{i_S j_{S^c}} p_{j_S i_{S^c}}.$$

We can then rewrite the recursion in the form

$$p_i' = p_i \frac{F_i(p)}{F(p)} - \frac{1}{F(p)} \sum_S c_S \sum_j D_{i,j}^S(p) w_{i,j}. \tag{2.17}$$

The sum is over all subsets $S \subset \{1, \ldots, k\}$, although it can be restricted to nontrivial subsets, as $D_{i,j}^S = 0$ for S the empty or the full set.

The numbers $D_{i,j}^S(p)$ are measures of "linkage disequilibrium" between the sets of loci S and loci S^c. They all vanish if and only if the population is in linkage equilibrium.

2.18 Lemma. *We have $D^S_{i,j}(p) = 0$ for every $S \subset \{1, 2, \ldots, k\}$ if and only if $p_{i_1 i_2 \ldots i_k} = p_{i_1} p_{i_2} \cdots p_{i_k}$ for every $i_1 i_2 \cdots i_k$.*

Proof. If the probabilities factorize, then it is immediate that $p_i p_j$ and $p_{i_S j_{S^c}} p_{j_S i_{S^c}}$ are the same products of marginal probabilities, and hence $D^S_{i,j}(p) = 0$, for every S. Conversely, we have $\sum_j D^S_{i,j}(p) = p_i - p_{i_S} p_{i_{S^c}}$, and hence $p_i = p_{i_S} p_{i_{S^c}}$ for every S if all $D^S_{i,j}(p)$ vanish. By summing this over the coordinates not in $S \cup T$ we see that $p_{i_S i_T} = p_{i_S} p_{i_T}$ for any disjoint subsets S and T of $\{1, \ldots, k\}$. This readily implies the factorization of p_i. ∎

It follows that the recursion (2.17) simplifies to the recursion (2.8) of the one-locus situation if the initial population is in linkage equilibrium. However, the set of all vectors p in linkage equilibrium is not necessarily invariant under the iteration (2.17), and in general the iterations may move the population away from linkage equilibrium. Depending on the fitness matrices many types of behaviour are possible. Even with as few as two loci the relative frequencies may show cyclic behaviour rather than stabilize to a limit. Also the fitness of the population may not increase, not even if the relative frequencies do converge and are close to their equilibrium point. We illustrate this below by a number of special cases. The failure of the increase in fitness is due to recombination, which creates new haplotypes without regard to fitness. It arises only if there is interaction (epistasis) between the loci.

That the population may not tend to or remain in "linkage equilibrium" is painful for the latter terminology. It should be remembered that "linkage equilibrium" received its name from consideration of dynamics without selection, where it is a true equilibrium. Some authors have suggested different names, such as "lack of association". Interestingly, the commonly used negation "linkage disequilibrium" has also been criticized for being misleading, for a different reason (see Section 9.1).

Marginal Frequencies. The marginal relative frequencies in the children's population can be obtained by summing over equation (2.17). Alternatively, for a subset $S \subset K = \{1, \ldots, k\}$ we can first obtain the relative frequency of a parent (A_{i_S}, A_{j_S}) as (with more lazy notation),

$$p_{i_S, j_S} = \sum_{g_{K-S}} \sum_{h_{K-S}} p_{i_S g_{K-S}} p_{j_S h_{K-S}} w_{i_S g_{K-S}, j_S h_{K-S}}.$$

Next, by the same argument as before we obtain that for, $T \subset K$,

$$\begin{aligned} p'_{i_T} &= \sum_{S \subset T} \sum_{j_{T-S}} \sum_{j_S} c^T_S p_{i_S j_{T-S}, j_S i_{T-S}} \\ &= \sum_{S \subset T} c^T_S \sum_{j_{K-S}} \sum_{h_{S \cup (K-T)}} p_{i_S j_{K-S}} p_{i_{T-S} h_{S \cup (K-T)}} w_{i_S j_{K-S}, i_{T-S} h_{S \cup (K-T)}}. \end{aligned}$$

Here c^T_S is the probability that there is a recombination between every endpoint of a segment created by S or S^c within the set of loci T.

In particular, the allele relative frequencies are obtained by choosing T a single locus $T = \{t\}$. Then there are two terms in the outer sum, corresponding to $S = \emptyset$ and $S = \{t\}$, with coefficients c_S^T both equal to $\frac{1}{2}$. The resulting formula can be written as

$$p'_{i_s} = \sum_{i:i_s=i_s} \sum_j p_i p_j w_{i,j} = p_{i_s} \sum_{j_s} p_{j_s} \bar{w}_{i_s,j_s}(p), \tag{2.19}$$

for $\bar{w}_{i_s,j_s}(p)$ given by

$$\bar{w}_{i_s,j_s}(p) = \sum_{i:i_s=i_s} \sum_{j:j_s=j_s} \frac{p_i}{p_{i_s}} \frac{p_j}{p_{j_s}} w_{i,j}.$$

This puts the recursion in the form (2.8), with the single-locus fitness taken to be $\bar{w}_{i_s,j_s}(p)$. The latter can be interpreted as the average fitness of allele pair (A_{i_s}, A_{j_s}) in the population, the product $(p_i/p_{i_s})(p_j/p_{j_s})$ being the conditional probability of a random individual being (A_i, A_j), and having fitness $w_{i,j}$, given that the individual's alleles at locus s are (A_{i_s}, A_{j_s}). However, an important difference with the one-locus situation is that the average fitness $\bar{w}_{i_s,j_s}(p)$ is dependent on p, and changes from generation to generation. Thus the marginal frequencies do not form an autonomous system, and their evolution does depend on the positions of the loci on the genetic map.

The recursions for higher order marginals T can similarly be interpreted as having the form (2.17) applied to T, with an average fitness.

Two Loci. In the case of two loci ($k = 2$) there are four different subsets S in the sum in (2.17). The trivial subsets $S = \emptyset$ and $S = \{1, 2\}$ contribute nothing, and the nontrivial subsets $S = \{1\}$ and $S = \{2\}$ are each other's complement and therefore have the same $D_{i,j}^S(p)$-value, equal to

$$D_{i_1i_2,j_1j_2}(p) = p_{i_1i_2} p_{j_1j_2} - p_{i_1j_2} p_{j_1i_2}.$$

The sum $\theta := c_{\{1\}} + c_{\{2\}}$ is equal to the recombination fraction between the two loci. Formula (2.17) therefore simplifies to

$$p'_i = p_i \frac{F_i(p)}{F(p)} - \frac{1}{F(p)} \theta \sum_j D_{i,j}(p) w_{i,j}. \tag{2.20}$$

Here we have assumed that the fitness $w_{i_1i_2,j_1j_2}$ depends on the two unordered sets $\{i_1, j_1\}$ and $\{i_2, j_2\}$ only.

The formula simplifies further in the case of two bi-allelic loci. Most of the coefficients $D_{i,j}$ are then automatically zero, and the four nonzero ones are plus or minus $D(p) := p_{11} - p_1 p_2$ each other, as shown in Table 2.1. Moreover, under the assumed symmetries the fitness values corresponding to the four combinations (i, j) of haplotypes with nonzero $D_{i,j}$ are identical. With the common value $w_{11,22} =$

$w_{12,21} = w_{21,12} = w_{22,11}$ denoted by w, the iterations become, for $11, 12, 21, 22$ the four haplotypes,

$$\begin{aligned} p'_{11} &\propto p_{11}F_{11}(p) - \theta D(p)w, \\ p'_{12} &\propto p_{12}F_{12}(p) + \theta D(p)w, \\ p'_{21} &\propto p_{21}F_{21}(p) + \theta D(p)w, \\ p'_{22} &\propto p_{22}F_{22}(p) - \theta D(p)w. \end{aligned} \tag{2.21}$$

The proportionality constant is the fitness $F(p)$.

Because an internal point $p = (p_{11}, p_{21}, p_{12}, p_{22})^T$ of (local) maximum fitness is necessarily a stationary point of the Lagrangian of F under the side condition $p^T 1 = 1$, it must satisfy $F_{ij}(p) \propto 1$ and hence $F_{ij}(p) = F(p)$, for every $ij = 11, 12, 21, 22$. The recursion formulas (2.21) show that such a vector p is also a fixed point of the iterations if and only if $D(p) = 0$. However, the four equations $F_{ij}(p) = 1$ form a system of linear equations determined by the fitness matrix $(w_{i,j})$ and there is no reason that a solution would satisfy $D(p) = 0$. Therefore, in general the extrema of the fitness function F do not coincide with the fixed points of (2.21). This suggests that the iterations may not necessarily increase the fitness, and this can indeed be seen to be the case in examples.

i/j	11	12	21	22
11	0	0	0	D
12	0	0	$-D$	0
21	0	$-D$	0	0
22	D	0	0	0

Table 2.1. Values of $D_{i,j}$ for two biallelic loci. The four possible haplotypes are labelled $11, 12, 21, 22$ and $D = p_{11} - p_1 p_2$.

2.22 EXERCISE. Prove the validity of Table 2.1. [Hint: for the anti-diagonal reduce D and a value such as $p_{11}p_{22} - p_{12}p_{21}$ both to a function of three of the four haplotype probabilities.]

No Selection. Choosing $w_{i,j} = 1$ for every (i, j), we regain the dynamics without selection considered in Sections 2.2 and 2.3. Then $F(p) = F_i(p) = 1$ for every i, and the sum on j in (2.17) can be performed to give

$$p'_i = p_i - \sum_S c_S (p_i - p_{i_S} p_{i_{S^c}}). \tag{2.23}$$

In the case that there are only two loci, the sum over S can be reduced to a single term, with leading coefficient θ, and we can regain the iteration expressed in Lemma 2.5. The next lemma generalizes the conclusion of this lemma to multiple loci. The condition that $c_S > 0$ for every S expresses that no subset of loci should be completely linked.

2.24 Lemma. *If $c_S > 0$ for every S and $w_{i,j} = 1$ for every (i,j), then the haplotype relative frequencies $p_i^{(n)}$ after n generations satisfy $p_i^{(n)} - p_{i_1}p_{i_2}\cdots p_{i_k} \to 0$ as $n \to \infty$, for $p_{i_1}, p_{i_2}, \ldots, p_{i_k}$ the marginal relative frequencies in the initial population.*

Proof. The recursion (2.23) implies that, for every i and every subset $T \subset \{1,\ldots,k\}$.

$$p_i' - p_{i_T}p_{i_{T^c}} = (1 - c_T - c_{T^c})(p_i - p_{i_T}p_{i_{T^c}}) - \sum_{S \neq T, T^c} c_S(p_i - p_{i_S}p_{i_{S^c}}).$$

Let Θ be a collection of subsets of $\{1,\ldots,k\}$ that contains for every nontrivial (i.e. not empty or the whole set) subset $T \subset \{1,\ldots,k\}$ either T or its complement T^c. Define $\mathbb{P}$ as the $(\#\Theta \times \#T)$-matrix whose rows and columns correspond to the elements of Θ, in arbitrary order, and whose Tth column has the number $2c_T$ in each row. Form vectors $p_i 1$ and $(p_{i_T}p_{i_{T^c}})$ with coordinates indexed by Θ with as Tth elements p_i and $p_{i_T}p_{i_{T^c}}$, respectively. The preceding display, for every $T \in \Theta$, can then be written

$$p_i' 1 - (p_{i_T}p_{i_{T^c}}) = (I - \mathbb{P})\big(p_i 1 - (p_{i_T}p_{i_{T^c}})\big).$$

Therefore, using similar notation for the relative frequencies in the consecutive generations, we infer that

$$\begin{aligned} p_i^{(n+1)}1 - \prod_s p_{i_s}1 &= (I - \mathbb{P})\Big(p_i^{(n+1)}1 - \prod_s p_{i_s}1\Big) + (p_{i_T}^{(n)}p_{i_{T^c}}^{(n)}) - \prod_s p_{i_s}1 \\ &= (I - \mathbb{P})^{n+1}\Big(p_i^{(0)}1 - \prod_s p_{i_s}1\Big) + \sum_{k=0}^{n}(I - \mathbb{P})^k\Big((p_{i_T}^{(n-k)}p_{i_{T^c}}^{(n-k)} - \prod_s p_{i_s}1\Big). \end{aligned}$$

The matrix $\mathbb{P}$ is a strictly positive stochastic matrix, and hence by the Perron-Frobenius theorem its eigenvalues have modulus smaller than 1. It follows that the spectral radius of the matrix $I - \mathbb{P}$ is strictly smaller than one, and hence $\|(I - \mathbb{P})^n\| \to 0$ as $n \to \infty$. Therefore the first term on the right side tends to zero. We also have that $\|(I - P)^n\| \leq Cc^n$ for some $c < 1$ and $C > 0$. (In fact $\|(I - P)^n\|^{1/n}$ tends to the spectral radius and hence there exists $d < 1$ such that $\|(I-P)^k\| < d^k$ for sufficiently large k. Then $\|(I-P)^n\| \leq C\|(I-P)^k\|^{n/k} \leq Cd^{n/k}$.) It follows that the terms of the sum are dominated by a constant times c^k. We can now proceed by induction on the number of loci. The terms in the sum refer to haplotype frequencies of haplotypes at the loci in T, which are smaller than k. Under the induction hypothesis the terms of the sum tend to zero as $n \to \infty$, for every fixed k. By the dominated convergence theorem the sum then also tends to zero. ∎

Additive Fitness. The fitness is said to be additive in the loci if, for marginal fitness measures $w_{s|i_s,j_s}$,

$$w_{i_1\cdots i_k, j_1\cdots j_k} = w_{1|i_1,j_1} + \cdots + w_{k|i_k,j_k}. \qquad (2.25)$$

In this situation the dynamics are much as in the case of a single locus: the population fitness never decreases and the iterations converge. Moreover, as in the case of no selection the population converges to linkage equilibrium.

The reason is that the fitness of the population depends only on the marginal allele frequencies, whose single-step iterations in turn do not dependent on the recombination probabilities c_S.

2.26 Theorem. *If $w_{i,j}$ satisfies (2.25) for functions $w_{s|i_s,j_s}$ that are symmetric in i_s, j_S, then the iterations (2.17) satisfy $F(p') \geq F(p)$ with equality only if $p = p'$. Furthermore, every sequence of iterates $p, p', p'', \ldots$ converges to a limit, where the only possible limit points are vectors $p = (p_i)$ of the form $p_i = p_{i_1} \cdots p_{i_k}$ that are also fixed points of (2.17) in the situation that the loci are completely linked.*

Proof. The fitness takes the form

$$F(p) = \sum_{s=1}^{k} \sum_{i_s} \sum_{j_s} p_{i_s} p_{j_s} w_{s|i_s,j_s},$$

This expression depends on p only through the marginal relative frequencies p_{i_s}. By (2.19) the latter satisfy a one-generation evolution that depends on the current joint relative frequency p, but not on the probabilities c_S. It follows that starting from p, the fitness in the next generation is the same no matter the map position of the loci. If the take the loci completely linked, then we can think of the haplotypes as single-locus alleles, where the number of possible alleles is equal to the set of possible vectors $i = i_1 \cdots i_k$, having "allele" relative frequencies p_i. Theorem 2.10 shows that the fitness increases unless $p = p'$.

For the proof of the last assertion see Lyubich, 9.6.13 and 9.6.11. ■

Multiplicative Fitness.

Cyclic Behaviour. Even in the two-locus, biallelic system (2.21) complex behaviour is possible. Lyubich (1992, 9.6.16-5) gives a theoretical example of a multiplicative fitness matrix, where the sequences of odd and even numbered iterations converge to different limits. (The recombination fraction in this example is bigger than 3/4, so the example devoid of genetical significance.) In Figure 2.1 we produce a more dramatic numerical example of the two-locus system, showing cyclic behaviour. The fitness matrix (with the haplotypes ordered as 11, 12, 21, 22) in this example is

$$\begin{pmatrix} 0.8336687 & 0.6606954 & 0.5092045 & 1.00000 \\ 0.6606954 & 1.3306800 & 1.0000000 & 0.22458 \\ 0.5092045 & 1.0000000 & 0.8072400 & 0.46357 \\ 1.0000000 & 0.2245800 & 0.4635700 & 1.41881 \end{pmatrix}.$$

The ouput of the system is sensitive to the starting point, which in the numerical simulation was chosen close to an (unstable) fixed point of the dynamical system.

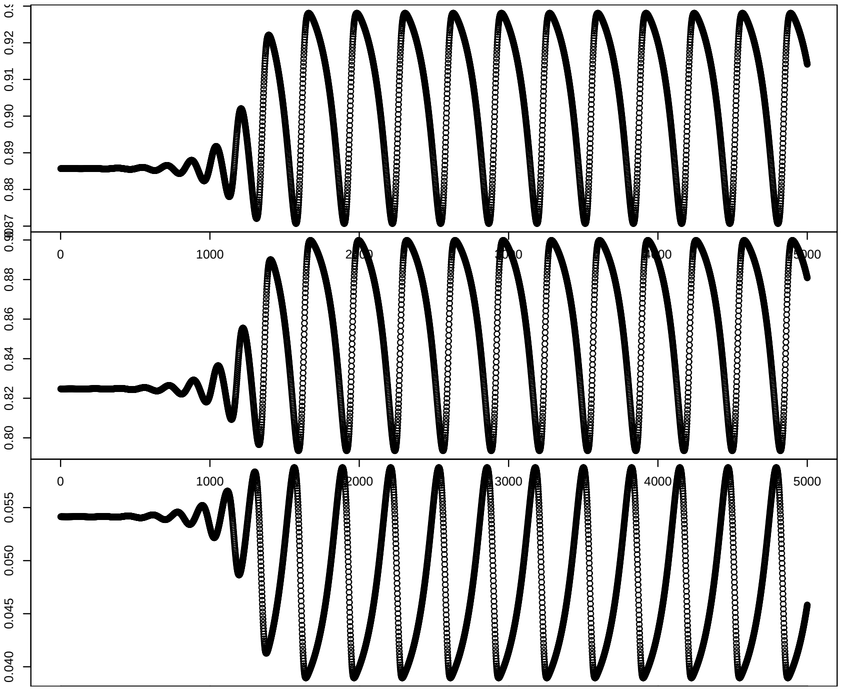

Figure 2.1. Example of cyclic behaviour in the two-locus, biallelic system (2.21). The top two panels give the allele relative frequencies at the two lcoi, and the bottom panel the linkage disequilibrium, each of 5000 generations. The fitness matrix is given in the text, and the starting vector is $p = (0.78460397, 0.10108603, 0.04013603, 0.07417397)$, corresponding to allele frequencies 0.88569 and 0.82474, and linkage disequilibrium $D = 0.05414$. [Source: Alan Hastings (1981), Stable cycling in discrete-time genetic models, Proc. Natl. Acad. Sci.78(11), 7224–7225.]

2.6.3 Fisher's Fundamental Theorem

Fisher's fundamental equation of natural selection relates the change in fitness of a population to the (additive) variance of fitness. We start with a simple lemma for the evolution at a single locus. Let $F(p)$ be the fitness of a population, $F_i(p)$ the marginal fitness of allele A_i, and $\sigma_A^2(p) = 2\sum_i p_i\big(F_i(p) - F(p)\big)^2$ twice the variance of the marginal fitness.

2.27 Lemma (Fisher's fundamental theorem). *Under the single-locus recursion (2.8),*

$$F(p') - F(p) = \frac{\sigma_A^2(p)}{F(p)} + F(p' - p).$$

Proof. For W the matrix $(w_{i,j})$ we can write $F(p') - F(p) = 2(p' - p)^T Wp + F(p' - p)$. Here we replace the first occurrence of $p' - p$ by the right side of equation (2.9), and note that

$$2\big(\mathrm{diag}\,(p)Wp - p\,F(p)\big)^T Wp = 2\sum_i \big(F_i(p) - F(p)\big)p_i F_i(p).$$

The right side is the variance $\sigma_A^2(p)$ of the marginal fitnesses, the fitness $F(p)$ being their mean. ∎

The term $F(p'-p)$ is negligible in the limit if $\max_{i,j}|w_{i,j}-1|\to 0$, a situation known as *weak selection.* (See Problem 2.28.) Fisher's fundamental theorem is therefore often quoted as an approximation in the (somewhat cryptic) form

$$\Delta\bar{w}\approx\frac{\sigma_A^2}{\bar{w}}.$$

Apparently Fisher considered this formula as fundamental to understanding evolution. He summarized it in the form of a law as: *the rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time.*[†]

2.28 EXERCISE. Suppose that $\|w-1\|:=\max_{i,j}|w_{i,j}-1|$ tends to zero. Show that $\sigma_A^2(p)=O\bigl(\|w-1\|^2\bigr)$ and $F(p'-p)=O\bigl(\|w-1\|^3\bigr)$. [Hint: if $w_{i,j}=1$ for every (i,j), then $F_i(p)=F(p)=1$ for every p and $F(p'-p)=0$. Deduce that $p'-p=O\bigl(\|w-1\|\bigr)$.]

Because Fisher played such an important role in the development of statistical genetics, there has been speculation about the exact meaning and formulation of his fundamental theorem. First it must be noted that Fisher interpreted the quantity $\sigma_A^2(p)$ not as twice the variance of marginal fitness, but as the *additive variance* of fitness viewed as a trait in the population. Consider a population of individuals (A_i,A_j) in Hardy-Weinberg equilibrium characterized by the allele relative frequency vector p. If (G_P,G_M) is the genotype of an individual chosen at random from the population and W is his fitness, then the best approximation to W by a random variable of the form $g(G_P)+g(G_M)$ (for an arbitrary function g, in the mean square sense) is the random variable $\mathrm{E}_p(W|\,G_P)+\mathrm{E}_p(W|\,G_M)-\mathrm{E}_pW$ (cf. Section 6.1.1). From the explicit formula $\mathrm{E}_p(W|\,G_P=A_i)=\sum_j p_jw_{i,j}=F_i(p)$, it follows that $\sigma_A^2(p)$ is the variance of this approximation.

The fundamental theorem itself can also be formulated in terms of "additive fitness", and some authors claim that Fisher meant it this way. The additive approximation $\mathrm{E}_p(W|\,G_P)+\mathrm{E}_p(W|\,G_M)-\mathrm{E}_pW$ to the fitness W could be taken as defining a new "additive fitness" measure of individual (A_i,A_j) as

$$\tilde{w}_{i,j}(p)=\mathrm{E}_p(W|\,G_P=A_i)+\mathrm{E}_p(W|\,G_M=A_j)-\mathrm{E}_pW.$$

This approximation depends on the relative frequencies p, because these determine the joint distribution of (G_P,G_M,W). We fix this at the current value p, but replace the relative frequencies p of the individuals in the population by p'. The change in "additive fitness" is then

$$\sum_i\sum_j p_i'p_j'\tilde{w}_{i,j}(p)-\sum_i\sum_j p_ip_j\tilde{w}_{i,j}(p)=\frac{\sigma_A^2(p)}{F(p)}. \tag{2.29}$$

[†] R.A. Fisher (1958), The genetical theory of natural selection, p 37.

The last equality follows by simple algebra (or see below). The point of the calculation is that this formula is exact, although one may question the significance of the expression on the left, which is a “partial change of partial fitness”.

Another point is that in this form the fundamental theorem extends to selection in multi-locus systems. If (G_P, G_M) is the pair of k-locus haplotypes of an individual, then we can in the same spirit as before define a best additive approximation of the fitness W as the projection of W, relative to mean square error computed under the population relative frequencies p, onto the space of all random variables of the form $\mathrm{E}_pW + \sum_s \bigl(g_s(G_{P,s}) + g_s(G_{M,s})\bigr)$, for $G_P = (G_{P,1}, \ldots, G_{P,k})$ and $G_M = (G_{M,1}, \ldots, G_{M,k})$ and g_s arbitrary functions. If this projection is given by functions $f_{s,p}$, then the *additive fitness* of an individual with haplotype pair (A_i, A_j) is defined as

$$\tilde{w}_{i,j}(p) = \mathrm{E}_pW + \sum_{s=1}^{k} \bigl(f_{s,p}(A_{i_s}) + f_{s,p}(A_{j_s})\bigr). \qquad (2.30)$$

This fitness is additive also in the sense of (2.25), but in addition the marginal fitnesses $f_{s,p}(A_{i_s}) + f_{s,p}(A_{j_s})$ are special. They depend on the current relative frequency vector p, but are fixed in evaluating the change in fitness to the next iteration.

2.31 Theorem (Fisher's fundamental theorem). *Under the multi-locus recursion (2.17), equation (2.29) is valid for the additive fitness defined in (2.30), with $\sigma_A^2(p)$ taken equal to the variance of the orthogonal projection in $L_2(p)$ of W onto the set of random variables of the form $\mathrm{E}_pW + \sum_s \bigl(g_s(G_{P,s}) + g_s(G_{M,s})\bigr)$.*

Proof. If $\Pi_{P,p}W$ and $\Pi_{M,p}W$ are the $L_2(p)$-projections of $W - \mathrm{E}_pW$ onto the spaces of mean zero random variables of the forms $\sum_s g_s(G_{P,s})$ and $\sum_s g_s(G_{M,s})$, respectively, then the left side of (2.29) is equal to the difference $\mathrm{E}_{p'}(\Pi_{P,p}W + \Pi_{M,p}W) - \mathrm{E}_p(\Pi_{P,p}W + \Pi_{M,p}W)$. By symmetry this is twice the paternal contribution.

Because the variable $\Pi_{P,p}W$ is a sum over the k loci, the expectations $\mathrm{E}_{p'}\Pi_{P,p}W$ and $\mathrm{E}_p\Pi_{P,p}W$ depend on $p' = (p'_i)$ and $p = (p_i)$ through the marginal frequencies $p'_{i_s} = \sum_{j:j_s=i_s} p'_j$ and $p_{i_s} = \sum_{j:j_s=i_s} p_j$ for the loci $s = 1, \ldots, k$ only. In fact,

$$\mathrm{E}_{p'}\Pi_{P,p}W - \mathrm{E}_p\Pi_{P,p}W = \sum_s \mathrm{E}_{p'}f_{s,p}(G_{P,s}) - \mathrm{E}_pf_{s,p}(G_{P,s}) = \sum_s \sum_{i_s} f_{s,p}(A_{i_s})\Bigl(\frac{p'_{i_s}}{p_{i_s}} - 1\Bigr)p_{i_s}.$$

By (2.19) the recursions for the marginal frequencies can be written as $p'_{i_s} = p_{i_s}\mathrm{E}_p(W|\, G_{P,i_s} = A_{i_s})$. It follows that the preceding display can be rewritten as

$$\sum_s \mathrm{E}_pf_{s,p}(G_{P,s})\bigl(\mathrm{E}_p(W|\, G_{P,s}) - 1\bigr) = \sum_s \mathrm{E}_pf_{s,p}(G_{P,s})W = \mathrm{E}_p(\Pi_{P,p}W)W.$$

Because $\Pi_{P,p}W$ is an orthogonal projection of W, the right side is equal to $\mathrm{E}_p(\Pi_{P,p}W)^2$. ■

The preceding theorem does not assume linkage equilibrium, but is valid for general multi-locus allele frequency vectors p. Inspection of the proof shows that dependence across loci is irrelevant, because of the assumed additivity of the fitness. As in all of this section, we have implicitly assumed independence of the paternal and maternal haplotypes (i.e. random mating). That assumption too is unnecessary, as the "additive fitness" is by definition also additive across the parental haplotypes. Thus Fisher's fundamental theorem becomes a very general result, in this form also known as an instance of *Price's theorem*, but perhaps by its generality suffers a bit in content.

Weak selection and epistasis??

* 2.7 Fertility Selection

Certain mating pairs may have more offspring than others. To model this, *fertility selection* attaches fitness weights to mating *pairs*, rather than to individuals. We still assume that mating pairs are formed at random, possibly after viability selection of the individuals, but a given mating pair $(A_i, A_j) \times (A_k, A_l)$ produces offspring proportional to *fertility weights* $f_{i,j\times k,l}$.

Viability and fertility selection can be combined by first selecting the individuals that enter mating pairs by viability weights $w_{i,j}$, and next applying fertility selection. This leads to the overall weight $w_{i,j}w_{k,l}f_{i,j\times k,l}$ for the mating pair $(A_i, A_j) \times (A_k, A_l)$. To simplify notation we can incorporate the viability weights into the fertility weights, and denote the overall weight by $w_{i,j\times k,l}$. On the other hand, if the fertility is multiplicative ($f_{i,j\times k,l} = \tilde{g}_{i,j}\tilde{g}_{k,l}$), then it is easier to incorporate the fertility weights in the viability weights, and fertility selection works exactly as viability selection. In the general case, fertility selection is more complicated to analyse.

Under fertility selection the two parents in a mating pair are not independent and hence zygotes are not composed of independent haplotypes. This makes it necessary to follow the successive populations (A_i, A_j) through their genotypic relative frequencies $p_{i,j}$. The mating pair $(A_i, A_j) \times (A_k, A_l)$ has relative frequency $p_{i,j}p_{k,l}w_{i,j\times k,l}$ in the population of all mating pairs and produces a child by independently recombining haplotypes A_i and A_j, and A_k and A_l.

2.7.1 Single Locus

For a child of type (A_i, A_j) the allele A_i is the father's paternal or maternal allele, and the allele A_j is the mother's paternal or maternal allele. The probability that in both cases it is the paternal allele is $1/4$, and the parent pair is then $(A_i, A_r) \times (A_j, A_s)$ for some r and s, which has relative frequency $p_{i,r}p_{j,s}w_{i,r\times j,s}$. This and the same observation for the other three cases shows that the relative frequency of a child of type (A_i, A_j) is

$$
\begin{aligned}
p'_{i,j} \propto \tfrac{1}{4}\sum_r\sum_s p_{i,r}p_{j,s}w_{i,r\times j,s} + \tfrac{1}{4}\sum_r\sum_s p_{i,r}p_{s,j}w_{i,r\times s,j} \\
+ \tfrac{1}{4}\sum_r\sum_s p_{r,i}p_{j,s}w_{r,i\times j,s} + \tfrac{1}{4}\sum_r\sum_s p_{r,i}p_{s,j}w_{r,i\times s,j}.
\end{aligned}
\tag{2.32}
$$

If the fertility weights are symmetric in the two parents ($w_{i,r\times j,s} = w_{j,s\times i,r}$ for every i, r, j, s), then the right side is symmetric in i and j. The relative frequencies of genotypes (A_i, A_j) and (A_j, A_i) are then equal after one round of offspring, and it is not much loss of generality to assume symmetry in the first generation. If $p_{i,j} = p_{j,i}$ for every i, j and moreover the fertilities are symmetric in the parents and depend on the unordered genotypes of the parents only ($w_{i,r\times j,s} = w_{r,i\times j,s}$ for every i, r, j, s), then the preceding display reduces to

$$p'_{i,j} \propto \sum_r\sum_s p_{i,r}p_{j,s}w_{i,r\times j,s}.$$

The proportionality factor is the average fertility $\sum_i\sum_j\sum_k\sum_l p_{i,j}p_{k,l}w_{i,j\times k,l}$ of a mating pair.

In general, the population will not be in Hardy-Weinberg equilibrium.

* 2.8 Assortative Mating

The assumption that individuals choose their mates purely by chance seems not very realistic. The situation that they are lead by phenotypes of their mates is is called *assortative mating* if they have a preference for mates with similar characteristics and *desassortative mating* in the opposite case.

We can model (des)assortative mating by reweighting mating pairs $(A_i, A_j) \times (A_r, A_s)$ by weights $w_{i,j\times r,s}$. Mathematically this leads to the exact situation of fertility selection. Therefore instead consider the more special situation of a sexe-dominated mating system, where a mating pair is formed by first choosing a father from the population, and next a mother of a genotype chosen according to a probability distribution dependent on the genotype of the father. If $p_{i,j}$ is the relative frequency of genotype (A_i, A_j) in the current population, then a mating pair $(A_i, A_j) \times (A_r, A_s)$ is formed with probability $p_{i,j}\pi_{r,s|i,j}$, for $(r, s) \to \pi_{r,s|i,j}$ given probability distributions over the set of genetic types (A_r, A_s). This gives the situation of fertility selection with weights of the special form $w_{i,j\times r,s} = \pi_{r,s|i,j}/p_{r,s}$. The fact that $\sum_{r,s}\pi_{r,s|i,j} = 1$, for every (i, j), creates a Markov structure that makes this special scheme easier to analyse.

2.8.1 Single Locus

The basic recursion is given by (2.32), but can be simplified to, for $\pi_{j|i,k} = \frac{1}{2}\sum_l(\pi_{j,l|i,k} + \pi_{l,j|i,k})$,

$$p'_{i,j} = \tfrac{1}{2}\sum_k(p_{i,k}\pi_{j|i,k} + p_{k,i}\pi_{j|k,i}).$$

In vector form, with p' and p the vectors with coordinates $p'_{i,j}$ and $p_{i,j}$, these equations can be written as $p' = A^T p$, for A the matrix with in its rsth row and ijth column the number

$$A_{rs,ij} = \pi_{j|r,s}(\tfrac{1}{2}1_{r=i\neq s} + \tfrac{1}{2}1_{r\neq i=s} + 1_{r=i=s}).$$

The sum over every row of A can be seen to be 1, so that A is a transition matrix of a Markov chain with state space the set of genotypes (A_i, A_j). The chain moves from state (A_r, A_s) to state (A_i, A_j) with probability $A_{rs,ij}$. These moves can be described in a more insightful way by saying that the chain determines its next state (A_i, A_j) by choosing A_i randomly from $\{A_r, A_s\}$ and by generating A_j from the probability distribution $\pi_{\cdot|r,s}$.

The dynamics $(p')^T = p^T A$ is described by considering the current relative frequency vector p as the current distribution of the Markov chain and p' as the distribution at the next time. Thus a current state (a "father") of type (A_r, A_s) is chosen according to the current distribution $p_{r,s}$, and the next state (a "child") is formed by choosing the child's "paternal allele" randomly from its father's alleles $\{A_r, A_s\}$ and its "maternal allele" according to the probability distribution $\pi_{\cdot|r,s}$.

The long term dynamics of a sequence of populations is given by the consecutive laws of the Markov chain. If the transition matrix is aperiodic and irreducible, then the sequence $p, p', p'', \ldots$ will converge exponentially fast to the unique stationary distribution of the transition matrix. A sufficient condition for this is that $\pi_{j|r,s} > 0$ for every j, r, s, meaning that no genotype (A_r, A_s) completely excludes individuals carrying an allele A_j. Of course, convergence will take place even if the transition matrix is reducible, as long as it is not periodic, but the limit will depend on the starting distribution. On the other hand, it is not too difficult to create periodicities,?? which will give cycling relative frequencies.

2.33 Example. Suppose that a father chooses a random mate with probability $1-\lambda$ and a mate of his own type otherwise. Because he would choose the mate based on phenotype and as usual we like to assume that genotypes (A_i, A_j) and (A_j, A_i) lead to the same phenotype, we understand the mate as an unordered genotype $\{A_i, A_j\}$. Thus a father (A_r, A_s) chooses a mate (A_i, A_j) with probability $\pi_{i,j|r,s} = (1-\lambda)p_{i,j} + \lambda q_{i,j}1_{\{i,j\}=\{r,s\}}$, for $q_{i,j} = p_{i,j}/(p_{i.j} + p_{j,i})$ the probability that a randomly chosen individual of type $\{A_i, A_j\}$ has ordered genotype (A_i, A_j). This leads to

$$\pi_{j|r,s} = (1-\lambda)\tfrac{1}{2}(p_{j.} + p_{.j}) + \lambda(1_{r=j=s} + \tfrac{1}{2}1_{r=j\neq s} + \tfrac{1}{2}1_{r\neq j=s}).$$

These probabilities do depend on the marginals of the current relative frequency vector p, which may change from generation to generation. The Markovian interpretation of the dynamics given previously is still valid, but the transition matrix changes every iteration. Convergence?? □

2.34 Example (Biallelic locus). If the transition probabilities $\pi_{j|r,s}$ are symmetric in r and s, then so are the transition probabilities $A_{rs.ij}$, and the corresponding Markov chain on the genotypes (A_i, A_j) can be collapsed into a Markov chain on the unordered genotypes $\{A_i, A_j\}$. For a biallelic locus with alleles a and A this gives a Markov chain on a state space of three elements, which we write as $\{aa, aA, AA\}$. The transition kernel is

$$\begin{pmatrix} \pi_{a|aa} & \pi_{A|aa} & 0 \\ \frac{1}{2}\pi_{a|Aa} & \frac{1}{2}\pi_{A|Aa} + \frac{1}{2}\pi_{a|Aa} & \frac{1}{2}\pi_{A|Aa} \\ 0 & \pi_{a|AA} & \pi_{A|AA} \end{pmatrix}$$

Apart from the zeros in the first and third row, the three rows of this matrix can be any probability vector. If the probabilities $\pi_{A|aa}$ and $\pi_{a|AA}$ are strictly between 0 and 1, then the chain possesses a unique stationary distribution, given by

$$(\pi_{aa}, \pi_{Aa}, \pi_{AA}) \propto \Bigl(\frac{\pi_{a|Aa}}{2\pi_{A|aa}}, 1, \frac{\pi_{A|Aa}}{2\pi_{a|AA}}\Bigr).$$

Thus all three types exist in the limit, but not necessarily in equal numbers, and their fractions may be far from the starting fractions. □

* 2.9 Mutation

In the biological model of meiosis as discussed so far pieces of parental chromosomes were recombined and then passed on without changes. In reality one or more base pairs may be substituted by other base pairs, some base pairs may be deleted, and new base pairs may be added (for instance by repetition). Any such change is referred to by the term *mutation*. Mutations play an important role in creating new genetic varieties, and are the drivers of evolutionary change. Because they are rare, they are typically excluded from consideration in genetic studies of pedigrees that involve only a few generations. The influence of mutation on evolution works over many generations.

Certain disease genes are thought to have arisen due to mutation and since then passed on to descendants. In Chapter?? we discuss approaches that try to find such genes by tracing such genes to their "common ancestor".

* 2.10 Inbreeding

Inbreeding is a deviation of random mating caused by a preference for relatives as partners, but the term is not very clearly defined. For instance, small populations may be considered to be inbred, even if within the population there is random mating.

In human genetics inbreeding is not often important. On the other hand, animals or plants may be inbred on purpose to facilitate genetic analysis or to boost certain phenotypes. In particular, there are several standard experimental designs to create strains of genetically identical individuals by inbreeding.

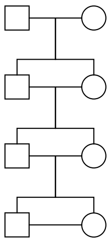

Figure 2.2. Four generations of inbreeding. A father and mother (the square and circle at the top) conceive a son and daughter who in turn are parents to a son and a daughter, etc.

The basis of these designs is to mate a son and daughter of a given pair of parents recursively, as illustrated for four generations in Figure 2.2. Eventually this will lead to offspring that is:

(i) identical at the autosomes.
(ii) homozygous at every locus.

The genetic make-up of the offspring depends on the genomes of the two parents and the realization of the random process producing the offspring.

2.35 Lemma. *Consider a sequence of populations each of two individuals, starting from two arbitrary parents and each new generation consisting of a son and a daughter of the two individuals in the preceding generation. In the absence of mutations there will arise with probability one a generation in which the two individuals are identical and are homozygous at every locus.*

Proof. Properties (i) and (ii) are retained at a given[1] locus in all future generations as soon as they are attained at some generation. Because there are only finitely many loci in the genome, it suffices to consider a single locus. The two parents have at most four different alleles at this locus, and the mating scheme can never introduce new allele types. If we label these four alleles by 1, 2, 3 and 4, then there are at most 16 different ordered genotypes $ab \in \{1,2,3,4\}^2$ for each individual and hence at most 256 different pairs *ab.cd* of ordered genotypes for the male and female in a given generation. The consecutive genotypic pairs form a Markov chain, whose transition probabilities can be computed with the help of Mendel's first law. The states 11.11, 22.22, 33.33 and 44.44 are clearly absorbing. A little thought will show that the Markov chain is irreducible and aperiodic with no other absorbing states. Absorption will happen eventually with probability one. ■

Suppose one repeats the inbreeding experiment as shown in Figure 2.2 with multiple sets of parents. This will yield a number of *strains* (biologists also speak of *models*) that each satisfy (i)–(ii), but differ from each other, depending on the parents and the breeding. The parents of the strains are often selected based on a phenotype of interest. If the parents differ strongly in their phenotype, then the descendants will likely also differ in their phenotype, and hence comparison of the strains will hopefully reveal the genetic determinants.

The next step in the experimental design is to cross the different strains. All descendants of a given pair of strains are identical, their two chromosomes being copies of the (single) chromosomes characterizing the strains. A *back-cross* experiment next mates a descendant with a parent of the strain, while an *intercross* mates two descendants of different strains??. The genomes of the resulting individuals are more regular than the genomes of randomly chosen individuals and can be characterized by markers measured on the parents of the strains. This facilitates statistical analysis.

3
Pedigree Likelihoods

A *pedigree* is a tree structure of family relationships, which may be annotated with genotypic and phenotypic information on the family members. In this chapter we show how to attach a likelihood to a pedigree, and apply it to *parametric linkage analysis*. The likelihood for observing markers and phenotypes of the individuals in a pedigree is written as a function of recombination fractions between markers and putative disease loci and "penetrances", and next these parameters are estimated or tested by standard statistical methods. The idea is that relative positions of markers, if not known, can be deduced from their patterns of cosegregation in families. Markers that are close are unlikely to be separated by crossovers, and hence should segregate together, and vice versa. A likelihood analysis permits to make this idea quantitative and form a "map" of the relative positions of marker loci. Similarly loci that are responsible for a disease can be positioned relative to this "genetic map" by studying the cosegregation of putative disease loci and marker loci and relating this to the disease phenotype. The estimation of recombination fractions between marker and disease loci is called "linkage analysis". With the help of a map function these recombination fractions can be translated into a "genetic map" giving the relative positions of the genes and markers.

3.1 Pedigrees

Figure 3.1 gives an example of a three-generation pedigree. Squares depict males and circles females, a horizontal line between a circle and a square means mating, and connections with a vertical component designate descendants. Squares and circles are filled or empty to indicate a binary phenotype of the individual; for ease of termininology "filled" will be referred to as "affected", and "empty" as "unaffected". Figure 3.1 shows eight individuals, who have been numbered arbitrarily by $1, 2, \ldots, 8$ for identification. Individuals 1 and 2 in the pedigree have children 3 and 4, and

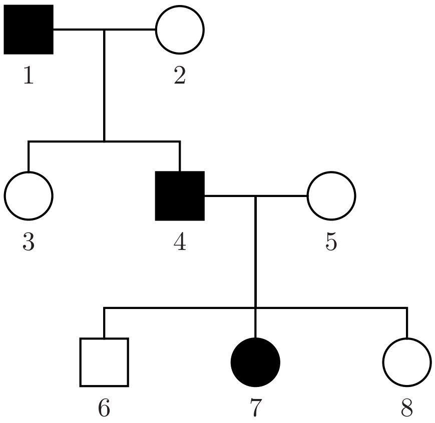

Figure 3.1. A three-generation pedigree.

individuals 4 and 5 have children 6, 7 and 8.

Individuals in a pedigree fall into two classes: founders and nonfounders. Individuals of whom at least one parent is also included in the pedigree are *nonfounders*, whereas individuals whose parents are not included are *founders*. This classification is specific to the pedigree and may run through the generations. For instance, in Figure 3.2 the grandparents 1 and 2, but also the spouse 5 in the second generation, are founders. The other individuals are nonfounders.

In this chapter we always assume that the founders of a pedigree are sampled independently from some given population. In contrast, given the founders the genetic make-up of the nonfounders is determined by their position in the pedigree and the chance processes that govern meiosis. The pedigree structure (who is having children with whom and how many) is considered given. Each nonfounder is the outcome of two meioses, a paternal and a maternal one. All meioses in the pedigree are assumed independent.

In typical pedigrees both parents of a nonfounder will be included, although it may happen that there is information on one parent only. In general we include all individuals in a pedigree that share a family relationship and on whom some information is available, and perhaps also both parents of every nonfounder. Individuals who share no known family relationship are included in different pedigrees. Our total set of information will consist of a collection of pedigrees annotated with the phenotypic and genotypic information.

Different pedigrees are always assumed to have been sampled independently from a population. This population may consist of all possible pedigrees, or of special pedigrees. If some pedigrees have more probability to be included than others, then this should be expressed through their likelihood. Statisticians speak in this case of “biased sampling”, geneticists of *ascertainment*. If a pedigree is included in the analysis, because one of the individual was selected first, after which the other individuals were ascertained, then this first individual is called the *proband*. A proband may have been selected because he of a disease he carries; the selection is then biased towards the disease.

3.2 Fully Informative Meioses

The pedigree in Figure 3.2 shows a family consisting of a father, a mother, a son and a daugher, and their alleles (labelled with the arbitrary names 1, 2, 3, or 4) at a given marker locus. The father and the daughter are affected, and we wish to investigate if the affection is linked to the marker locus. In practice we have more than one pedigree and/or a bigger pedigree, but this small pedigree serves to introduce the idea. The family is assumed to have been drawn at random from the population.

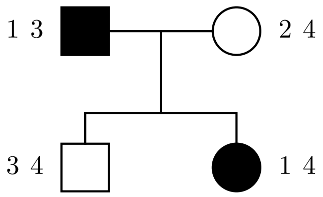

Figure 3.2. Pedigree showing a nuclear family, consisting of father, mother and two children, and their unordered genotypes at a marker location. The father and daughter are affected.

For further simplicity suppose that the affection is known to be caused by a single allele A at a single biallelic locus (a *Mendelian disease*), rare, and fully dominant without phenocopies. Dominance implies that an individual is affected if he has unordered genotype AA or Aa at the disease locus, where a is the other allele at the disease locus. The assumption that the affection is rare, makes both unordered genotypes Aa and AA rare, but the genotype Aa much more likely than the genotype AA (under Hardy-Weinberg equilibrium). Under the added assumption that no individual with genotype aa is affected ("no phenocopies"), the affection status indicated in Figure 3.2 makes it reasonable to assume that the unordered genotypes at marker and disease location are as in Figure 3.3.

Thus far we have considered unordered genotypes. The next step is to try and resolve the *phase* of the genotypes, i.e. to reconstruct the pairs of haplotypes for the two loci. The situation in Figure 3.3 is fortunate in that the marker alleles of the parents are all different. This allows to decide with certainty which allele the two parents have segregated to the children. In fact, the marker alleles of the two children, although considered to form unordered genotypes so far, have already been

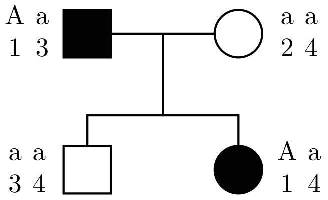

Figure 3.3. Pedigree showing a nuclear family, consisting of father, mother and two children, and their unordered genotypes at a marker location and the disease location. The disease is assumed to be Mendelian, and dominant with no phenocopies.

written in the "correct" paternal/maternal order. For instance, the allele 3 of the son clearly originates from the father and allele 4 from the mother. Reconstructing the phase requires that we put the alleles at the disease locus in the same order. The mother and the son are homozygous at the disease locus, and hence the ordering does not matter. For the daughter it is clear that both the disease allele A and the marker allele 1 were received from the father and hence her phase is known. Thus we infer the situation shown in Figure 3.4. The phase of the mother has been resolved in this figure, in the sense that her haplotypes are $a2$ and $a4$, even though the positioning of the haplotypes ($a2$ left, $a4$ right) is not meant to reflect parental or maternal origin, as this cannot be resolved.

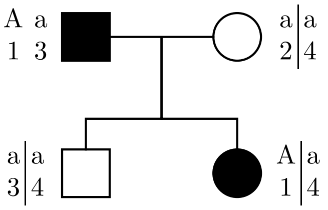

Figure 3.4. Pedigree showing a nuclear family, consisting of father, mother and two children, and their genotypes at a marker location and the disease location, including phase information for the mother and the children.

The phase of the father cannot be resolved from the information in the pedigree. If we care about haplotypes, but not about the (grand)paternal and (grand)maternal origins of the alleles, then there are clearly two possibilities, indicated in Figure 3.5.

There are four meioses incorporated in the pedigree: the father segregated a gamete to each of the children, and so did the mother. A meiosis is said to be *recombinant* if the haplotype (gamete) passed on by the parent consists of alleles taken from different parental chromosomes. Given the pedigree on the left in Figure 3.5, the father segregated the haplotype $a3$ to his son and the haplotype $A1$ to his daughter, and both are nonrecombinant. Given the pedigree on the right in Figure 3.5, the father segregated the same haplotypes, but both meioses were re-

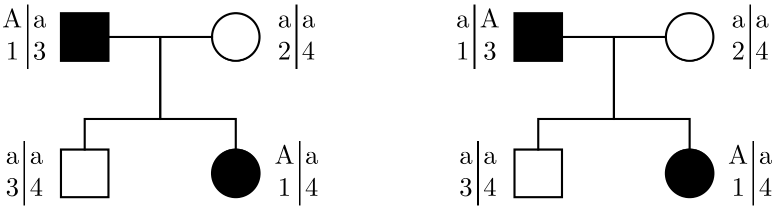

Figure 3.5. Pedigrees showing a nuclear family, consisting of father, mother and two children, and their ordered genotypes at a marker location and the disease location.

combinant. The mother segregated the haplotype $a4$ to both children, but neither for the pedigree on the left nor for the one on the right can the meioses be resolved to be recombinant or not. The four meioses are assumed to be independent, and hence independently recombinant or not. Recombination in a single meiosis occurs with probability equal to the recombination fraction between the disease and marker locus, which is directly related to their genetic map distance.

Under linkage equilibrium the two pedigrees in Figure 3.5 are equally likely. Because the left pedigree implies two nonrecombinant meioses and the right one two recombinant meioses, it is reasonable to assign the original pedigree of Figure 3.2 the likelihood

$$\tfrac{1}{2}(1-\theta)^2 + \tfrac{1}{2}\theta^2,$$

where θ is the recombination fraction between the disease and marker locus.

As a function of θ this function is decreasing on the interval $[0, \frac{1}{2}]$. Therefore, the maximum likelihood estimator for the recombination fraction is $\hat{\theta} = 0$, indicating that the disease locus is right on the marker locus. Of course, the small amount of data makes this estimate rather unreliable, but the derivation illustrates the procedure. In practice we would have a bigger pedigree or more than one pedigree. The total likelihood would be defined by multiplying the likelihoods of the individual pedigrees.

In practice rather than estimating the recombination fraction one often tests the null hypothesis $H_0: \theta = \frac{1}{2}$ that the recombination fraction is $\frac{1}{2}$, i.e. that there is no linkage between marker and disease. This can be done by the likelihood ratio test. In the example the log likelihood ratio statistic is

$$\log \frac{\frac{1}{2}(1-\hat{\theta})^2 + \frac{1}{2}\hat{\theta}^2}{\frac{1}{2}(1-\hat{\theta}_0)^2 + \frac{1}{2}\hat{\theta}_0^2},$$

with $\hat{\theta}$ the maximum likelihood estimator and $\hat{\theta}_0 = \frac{1}{2}$ the maximum likelihood estimator under the null hypothesis. The null hypothesis is rejected for large values of this statistic. In that case it is concluded that a gene in the "neighbourhood" of the marker is involved in causing the disease.

3.3 Pedigree Likelihoods

Consider a more formal approach based on writing a likelihood for the observed data, the pedigree in Figure 3.2. We assume that the father and mother are chosen independently at random from a population that is in combined Hardy-Weinberg and linkage equilibrium. Under this assumption the probabilities of observing the pedigrees with the indicated marker and disease genotypes on the left or on the right in Figure 3.5 are given by

$$p_A p_a p_1 p_3 \times p_a^2 p_2 p_4 \times \tfrac{1}{2}(1-\theta)\tfrac{1}{2} \times \tfrac{1}{2}(1-\theta)\tfrac{1}{2}, \tag{3.1}$$

and

$$p_A p_a p_1 p_3 \times p_a^2 p_2 p_4 \times \tfrac{1}{2}\theta\tfrac{1}{2} \times \tfrac{1}{2}\theta\tfrac{1}{2}, \tag{3.2}$$

respectively. Here p_A and p_a are the frequencies of alleles A and a at the disease locus, and p_1, p_2, p_3, p_4 are the frequencies of the marker alleles in the population. The probabilities are computed by first writing the probability of drawing two parents of the given types from the population and next multiplying this with the probabilities that the children have the indicated genotypes given the genotypes of their parents. The structure of the pedigree (father, mother and two children) is considered given. The probabilities of drawing the parents are determined by the assumption of combined Hardy-Weinberg and linkage equilibrium. The conditional probabilities of the genotypes of the children given the parents are determined according to Mendelian segregation and the definition of the recombination fraction θ. For instance, for the pedigree on the left $p_A p_a p_1 p_3$ is the probability that an arbitrary person has the genotype of the father, $p_a^2 p_2 p_4$ is the probability that an arbitrary person has the genotype of the mother, $\frac{1}{2}(1-\theta)\frac{1}{2}$ is the probability that the son has the given genotype given the parents (the father must choose to pass on allele a and then not recombine; the mother must choose allele 4) and $\frac{1}{2}(1-\theta)\frac{1}{2}$ is the probability that the daughter is as indicated given the parents' genotypes. We multiply the four probabilities, because the parents are founders of the pedigree and are assumed to have been selected at random from the population, while the four meioses are independent by assumption.

In the preceding section we argued that the two annotated pedigrees in Figure 3.5 are the only possible ones given the observed pedigree in Figure 3.2. As the two pedigrees have likelihoods as in the preceding displays and they seem equally plausible, it seems reasonable to attach the likelihood

$$\begin{aligned}&\tfrac{1}{2}\bigl[p_A p_a p_1 p_3 \times p_a^2 p_2 p_4 \times \tfrac{1}{2}(1-\theta)\tfrac{1}{2} \times \tfrac{1}{2}(1-\theta)\tfrac{1}{2}\bigr]\\ &\qquad + \tfrac{1}{2}\bigl[p_A p_a p_1 p_3 \times p_a^2 p_2 p_4 \times \tfrac{1}{2}\theta\tfrac{1}{2} \times \tfrac{1}{2}\theta\tfrac{1}{2}\bigr]\end{aligned} \tag{3.3}$$

to the annotated pedigree in Figure 3.2. This is not at all the likelihood $\frac{1}{2}(1-\theta)^2 + \frac{1}{2}\theta^2$ that was found in the preceding section. However, as functions of the recombination fraction θ the two likelihoods are proportional, and as likelihood functions they are equivalent.

We would like to derive a better motivation for combining the likelihoods for the two pedigrees in Figure 3.5 by taking their average. The key is that we would like to find the likelihood for the observed data, which is the annotated pedigree in Figure 3.2. Figure 3.5 was derived from Figure 3.2, but contains more information and consists of two pedigrees. Suppose we denote the observed data contained in the annotated pedigree in Figure 3.2 by x, a realization of a random variable X that gives the unordered marker genotypes of the four individuals in the pedigree. We would like to find the likelihood based on observing X, which is the density $p_\theta(x)$ viewed as function of the parameter θ. The annotated pedigrees in Figure 3.5 contain more information, say (x, y), where y includes the genotypes at the disease locus and the phase information. The two pedigrees in Figure 3.5 correspond to (x, y) for the same observed value of x, but for two different possible values of y. The probabilities in (3.1) and (3.2) are exactly the density $q_\theta(x, y)$ at the two possible realized values (x, y) of the vector (X, Y) that gives the ordered genotypes at both marker and disease locus together with phase information. Since we only observe X, the density of (X, Y) does not give the appropriate likelihood. However, the density of X can be derived from the density of (X, Y) through marginalization:

$$p_\theta(x) = \sum_y p_\theta(x, y).$$

The sum is over all possible values of y. In the preceding section it was argued that only two values of y are possible for the realized value of x, and they are given in Figure 3.5. The likelihood based on the observed value x in Figure 3.2 is therefore the sum of the probabilities in (3.1) and (3.2). Thus we obtain (3.3), but without the factors $\frac{1}{2}$. These factors seemed intuitive ("each of the two possibilities in Figure 3.5 has probability $\frac{1}{2}$"), but it would be better to omit them. Of course, from the point of view of statistical inference multiplication of the likelihood by $\frac{1}{2}$ is inconsequential, and we need not bother. (You can probably also manage to interprete the likelihood with the $\frac{1}{2}$ as a "conditional likelihood" of some type.)

Actually in the preceding section we made life simple by assuming that each affected individual has disease genotype Aa and each unaffected individual has genotype aa. In reality things may be more complicated. Some diseased people may have genotype AA and some may even have genotype aa (*phenocopies*) and not every individual with AA or Aa may be ill (*incomplete penetrance*). Let f_{AA}, f_{Aa} and f_{aa} be the *penetrances* of the disease: individuals with unordered genotypes AA, Aa or aa are affected with probability f_{AA}, f_{Aa}, or f_{aa}. (As is often done, we assume that the penetrances do not depend on paternal and maternal origin of the alleles: $f_{Aa} = f_{aA}$ if Aa and aA are ordered genotypes.) So far we have assumed that $f_{AA} = 1 = f_{Aa}$ and $f_{aa} = 0$, but in general the penetrances may be strictly between 0 and 1. Then besides the two pedigrees in Figure 3.5 many more pedigrees will be compatible with the observed data in Figure 3.2. To find the likelihood for the observed data we could enumerate all possibilities, write the probability of each possibility, and add all these expressions. We have the same observed data x, but many more possible values of y, and find the likelihood for observing X by the same method as before.

Even for the simple pedigree in Figure 3.2 enumerating all possibilities can be forbidding. In particular, if all three penetrances f_{AA}, f_{Aa}, or f_{aa} are strictly between 0 and 1, then every of the four individuals may have disease genotype AA, Aa or aa, irrespective of affection status. The disease genotypes given in Figure 3.3 are by far the most likely ones if allele A is rare and $f_{aa} \approx 0$, but correct inference requires that we also take all other possibilities into account. There are 3^4 possible unordered disease genotypes for the set of four individuals in the pedigree of Figure 3.2, and for each of these there may be 1 to 2^3 possible resolutions of the phase of the genotypes. We need a computer to perform the calculations. For somewhat bigger pedigrees we even need a fast computer and good algorithms to perform the calculations in a reasonable time.

Conceptually, there is no difficulty in implementing this scheme. For instance, in Figure 3.2 the marker genotypes of the children can unequivocally be determined to be in the order as given (left paternal, right maternal). We do not care about paternal and maternal origins of the marker alleles of the parents. We can therefore enumerate all possible disease/marker genotypes, by adding the 4^4 ordered disease genotypes (4 individuals, each with genotype A, Aa, aA or aa), defining the phase by the order in which the alleles are written. Figure 3.5 gives two of the 256 possibilities. Given penetrances between 0 and 1 the probabilities of the two pedigrees in Figure 3.5 must be revised (from (3.1) and (3.2)) to

$$p_A p_a p_1 p_3 f_{Aa} \times p_a^2 p_2 p_4 (1 - f_{aa}) \times \tfrac{1}{2}(1-\theta)\tfrac{1}{2}(1 - f_{aa}) \times \tfrac{1}{2}(1-\theta)\tfrac{1}{2} f_{Aa},$$

and

$$p_A p_a p_1 p_3 f_{Aa} \times p_a^2 p_2 p_4 (1 - f_{aa}) \times \tfrac{1}{2}\theta\tfrac{1}{2}(1 - f_{aa}) \times \tfrac{1}{2}\theta\tfrac{1}{2} f_{Aa}.$$

These expressions must be added to the 254 other expressions of this type (possibly equal to 0) to obtain the likelihood for observing the annotated pedigree in Figure 3.2.

In deriving the preceding expressions it has been assumed that given their genotypes the individuals are independently affected or unaffected (with probabilities given by the penetrances f_{aa}, f_{Aa} and f_{AA}). This is a common assumption, which is not necessarily realistic in case the disease is also determined by environmental factors, which may be common to the individuals in the pedigree. The influence of the environment is particularly important for complex traits and is discussed further in Chapter 8.

Missing marker information can be incorporated in the likelihood by the same method of enumerating all possibilities. For instance, in Figure 3.6 the marker information on the mother is missing (so that there is no information about the mother at all). From the genotypes of the children it is clear that the mother must have at least one marker allele 4, but the other allele cannot be resolved. The likelihood for the pedigree can be constructed by considering the possibilities that the missing marker allele is of type 1, 3 or 4 or another known type for the locus, and adding their probabilities. Because the number of possibilities is unpleasantly large, a better algorithm than a complete listing is advisable. For instance, one possibility for the missing markers in Figure 3.6 is given by Figure 3.2 and we have seen that this leads to 256 possible annotated pedigrees with information on disease locus and phase. The other possibilities for the missing marker contribute comparable numbers of annotated pedigrees.

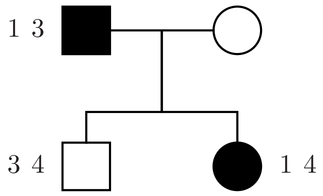

Figure 3.6. Pedigree showing a nuclear family, consisting of father, mother and two children, and the unordered genotypes of father and children at a marker location. The genotype of the mother is not observed. The father and daughter are affected.

3.3.1 Multilocus Analysis

Extension of the likelihood analysis to more than one marker and multigenic diseases is straightforward, except that the computational complications increase rapidly, and we need a workable model for joint recombination events.

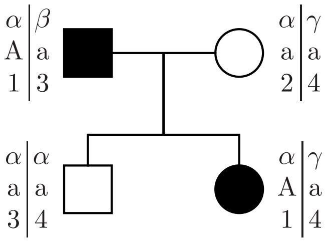

Figure 3.7. Pedigree showing a nuclear family, consisting of father, mother and two children, and their ordered genotypes at two marker locations and the disease location.

As an example consider the pedigree in Figure 3.7. It is identical to the pedigree in the left panel of Figure 3.5, except that marker information on an additional locus, placed on the other side of the disease locus has been added. The pedigree has been annotated with complete information on phase and disease locus alleles. Hence, in practice it will be only one of the many possible pedigrees that correspond to the observations, which would typically consist of only the unordered genotypes at the two marker loci. The likelihood would be the sum of the likelihood of the pedigree in Figure 3.7 and of the likelihoods of all the other possible pedigrees.

In Figure 3.7 the disease locus has been placed between the two marker loci. In the following we shall understand this to reflect its spatial position on the genome.

In practice, with only the marker loci given, the positioning would be unknown, and one would compute the likelihood separately for each possible positioning of the loci, hence also for the case that the disease locus is left or right (or rather "above" or "below" in the figure) of both marker loci.

Under the Poisson/Haldane model recombinations in disjoint intervals are independent. If we denote the recombination fractions of the interval between first marker and disease locus by θ_1 and between disease locus and second marker by θ_2, then the likelihood of the pedigree is

$$\begin{aligned}&p_\alpha p_\beta p_A p_a p_1 p_3 f_{Aa} \times p_\alpha p_\gamma p_a^2 p_2 p_4 (1-f_{aa})\\ &\qquad\times \tfrac{1}{2}\theta_1(1-\theta_2)\bigl[\tfrac{1}{2}\theta_1(1-\theta_2)+\tfrac{1}{2}(1-\theta_1)\theta_2\bigr](1-f_{aa})\\ &\qquad\times \tfrac{1}{2}(1-\theta_1)(1-\theta_2)\bigl[\tfrac{1}{2}\theta_1\theta_2+\tfrac{1}{2}(1-\theta_1)(1-\theta_2)\bigr]f_{Aa}.\end{aligned}\tag{3.4}$$

The appearance of the terms $\frac{1}{2}\theta_1(1-\theta_2)+\frac{1}{2}(1-\theta_1)\theta_2$ and $\frac{1}{2}\theta_1\theta_2+\frac{1}{2}(1-\theta_1)(1-\theta_2)$ (in square brackets) indicates a novel difficulty. Because the mother is homozygous at the disease locus, it is impossible to know whether she segregated her paternal or maternal disease allele a. Without this information it cannot be resolved whether recombination occurred between the loci and hence when writing down the likelihood we sum over the two possibilities. In general we can solve such ambiguities by annotating the pedigree for each locus both with the ordered founder genotypes and for each meiosis which of the two parental alleles is segregated. This information is captured in the "inheritance indicators" in Section 3.6.

The likelihood in the preceding display, and the observed likelihood of which the display gives one term, is a function of the two recombination fractions θ_1 and θ_2. For known marker loci, the genetic distance between the two markers is known, and hence the pair of parameters (θ_1,θ_2) can be reduced to a single parameter. (The known recombination fraction between the markers is equal to $\theta_{12}=\theta_1(1-\theta_2)+(1-\theta_1)\theta_1$; we can express one of θ_1 or θ_2 in θ_{12} and the other parameter.) The likelihood can next be maximized to estimate this parameter. Alternatively, geneticists usually test the hypothesis that the disease locus is at recombination fraction $\frac{1}{2}$ versus the hypothesis that the disease locus is at a putative locus between the markers using the likelihood ratio test, for every given putative locus (in practice often spaced 0.5 cM).

There is no conceptual difficulty in extending this analysis to include more than two marker loci. Incorporating more loci increases the power of the test, but also increases the computational burden. Markers with many alleles ("highly polymorphic markers") permit to resolve the paths by which the alleles are segregated and hence help to increase statistical power. If the disease locus is enclosed by two of such informative markers, then adding further markers outside this interval will not help much. In particular, under the Haldane model with marker loci $M_1,\ldots,M_k$, the recombination events between M_i and M_{i+1} are independent of the recombination events before M_i and past M_{i+1}. Thus if the disease locus is between M_i and M_{i+1} and the segregation at M_i and M_{i+1} can be completely resolved, then the markers before M_i and past M_{i+1} will not help in locating the disease locus. (These additional markers would just add a multiplicative factor to the likelihood.)

In practice, the phase of segregation at a given marker cannot be perfectly resolved and nearby markers may be helpful, but only to the extent of resolving segregation of the markers near the disease locus.

The likelihood (3.4) employs the Haldane map function for the joint probabilities of recombination or not, in the two intervals between first marker, disease locus and second marker. For a general map function these probabilities can be expressed in the map function as well. For instance, for m_1 and m_2 the genetic map distances of the two intervals, the probability of recombination in the first interval and no recombination in the second interval is given by (see Theorem 1.3)

$$\begin{aligned} P(R_1 = 1, R_2 = 0) &= \tfrac{1}{4}\bigl(P(N_1 + N_2 > 0) - P(N_1 > 0) - P(N_2 > 0)\bigr) \\ &= \tfrac{1}{2}\Bigl[\theta\bigl(\tfrac{1}{2}(m_1 + m_2)\bigr) + \theta\bigl(\tfrac{1}{2}m_1)\bigr) - \theta\bigl(\tfrac{1}{2}m_2\bigr)\Bigr]. \end{aligned}$$

The second equality follows from the definition (1.5) of map function. The right side of the display would replace the expression $\theta_1(1-\theta_2)$ in (3.4). The joint probabilities of recombinations in other pairs of adjacent intervals can be expressed similarly.

A multipoint analysis with more than two marker loci requires the joint probabilities of recombinations over three or more intervals. As noted in Section 1.3, in general a map function is not sufficient to express these probabilities, but other properties of the chiasmata process must be called upon. Theorem 1.3 shows how to express these probabilities in the avoidance probabilities of the chiasmata process. Of course, under the Haldane/Poisson model the occurrences of recombinations over the various intervals are independent and the map function suffices to express their probabilities.

The analysis can in principle also be extended to affections or traits that are caused by more than one gene. One simply places two or more loci among the markers and computes a likelihood for the resulting annotated pedigree, marginalizing over the unobserved genotypes at the "putative loci". The penetrances become functions of the genotypes at the vector of putative loci. Writing plausible models for the penetrances may be difficult, adding to the difficulty of the necessary computations.

3.3.2 Penetrances

If the measured phenotype is binary and caused by a single biallelic disease gene, then three penetrance parameters f_{AA}, f_{Aa} and f_{aa} suffice to specify the genetic model, for A and a the alleles at the disease locus. However, phenotypes may have many values (and may even be continuous variables), diseases may be multigenic, and disease genes may be multiallelic. Then a multitude of penetrance parameters is necessary to write a pedigree likelihood. These parameters will often not be known a-priori and must be estimated from the data.

Furthermore, penetrances could be made dependent on covariates, such as age or sex. Possible shared environmental influences on the phenotypes can also be incorporated. To keep the dimension of the parameter vector under control, covariates are often discretized: a population is divided up in subclasses, and penetrances are assumed to be constant in subclasses.

3.3.3 Why Hardy-Weinberg and Linkage Equilibrium?

In the preceding the likelihoods of the founders were computed under the assumption that these are chosen independently from a population that is in Hardy-Weinberg and linkage equilibrium. Actually most of the work is in listing the various possible pedigrees and working out the probabilities of the meioses given the parents. The equilibrium assumptions play no role in this, and other models for the founder genotypes could easily be substituted.

Why are the equilibrium assumptions made? One reason is in the computation of the maximum of the likelihood. Under Hardy-Weinberg and linkage equilibrium the observed unordered genotypes of a founder contribute the same multiplicative factor to the likelihood of every possible fully annotated pedigree, given by the product of the population marginal frequencies of all the founder's single locus alleles. Consequently, these founder genotypes contribute the same multiplicative factor to the likelihood of the observed pedigree, as this is the sum over the likelihoods of fully annotated pedigrees. Because multiplicative factors are of no importance in a likelihood analysis, this means that this part of the founder probabilities drop out of the picture.

As an alternative model suppose that we would assume random mating, but not linkage equilibrium. If the founders are viewed as chosen from a given population, then their genotypes are random combinations of two haplotypes and would add factors of the type $h_i h_j$ to the likelihood, where $h_1, \ldots, h_k$ are the haplotype relative frequencies in the population. Even in a two-locus analysis of marker loci, these contributions would depend on the (unknown) phase of the unordered genotypes at the two loci and hence would not be the same for every possible fully annotated (phase-resolved) pedigree. Therefore, in a likelihood analysis the haplotype frequencies would not factor out, but remain inside the sum over the annotated pedigrees.

The preceding concerns observed marker genotypes, and unfortunately not to all founder genotypes. The relative frequencies of the the disease alleles will factor out of the likelihood only if in every possible annotation of the likelihood the (unordered) disease genotypes contain the same alleles. In the preceding this was true for the two annotations of the pedigree in Figure 3.2 given in Figure 3.3, but this is typically not the case for all possible annotated pedigrees without the assumptions of full penetrance and absence of phenocopies. A second problem arises if some marker genotypes are not observed as in Figure 3.6, which requires summing out over all possible missing genotypes.

Even in these cases the equilibrium assumptions simplify the expression of the likelihood, to an extent that given the present computing power is desirable.

Often there are also independent estimates of allele frequencies available, which are used to replace these quantities in the likelihood by numbers.

3.4 Parametric Linkage Analysis

In the preceding sections we have seen how to express the likelihood of an observed pedigree in penetrances and recombination fractions between marker and disease loci. If one or more recombination fractions are unknown, these can be viewed as parameters of a statistical model, resulting in an "ordinary" likelihood function for the observed pedigree. Statistical inference based on this is known as *parametric linkage analysis.*

The location of the maximum of the likelihood function is of course a reasonable estimator of the unknown parameters. However, one must keep in mind that the disease loci may not belong to the marker area under consideration. Geneticists therefore typically report their analysis in terms of a test of the null hypothesis that the disease locus is unlinked to the observed markers. If the test rejects, then the disease locus is estimated by the maximum likelihood estimator.

Under the assumption that there is at most a single causal locus in the marker area under consideration, the location of this locus can be parametrized by a single parameter θ. If this is chosen equal to a recombination fraction with a marker, then the null hypothesis $H_0\colon \theta = 1/2$ expresses that the disease is unlinked. It can be tested with the likelihood ratio test, which rejects for large values of the ratio

$$\frac{\ell(\hat\theta)}{\ell(\frac12)}.$$

Here $\theta \mapsto \ell(\theta)$ is the likelihood for the model, and $\hat\theta$ the maximum likelihood estimator.

Under mild conditions (e.g. that the number of informative meioses or the number of independently sampled pedigrees tends to infinity), we can employ asymptotic arguments to derive an approximation to the (null) distribution of (twice) the log likelihood ratio statistic, from which a critical value of p-value can be derived. In the present situation the asymptotic distribution of twice the log likelihood ratio statistic under the null hypothesis is slightly unusual, due to the fact that the null hypothesis corresponds to the boundary point $\theta = 1/2$ of the possible interval $[0, 1/2]$ of recombination fractions. It is not a standard chisquare distribution, but a $1/2$–$1/2$ mixture of the chisquare distribution with one degree of freedom and a pointmass at 0. (See Example 14.19.) The critical value is therefore chosen as the upper 2α-quantile of the chisquare distribution with one degree of freedom.

In genetics it is customary to replace the log likelihood ratio by the *LOD score* (from "log odds"), which is the log likelihood ratio statistic with the natural logarithm replaced by the logarithm at base 10. Because ${}^{10}\log x = {}^{10}\log e \, \log x$ and ${}^{10}\log e \approx 0.434$, a LOD score is approximately 0.434 times the log likelihood ratio statistic. In practice a LOD-score of higher than 3 is considered sufficient proof of linkage. This critical value corresponds to a p-value of 10^{-4}, and is deliberately, even though somewhat arbitrarily, chosen smaller than usual in the light that one often carries out the test on multiple chromosomes or marker areas simultaneously.

The LOD-scores are typically reported in the form of a graph as a function of the position of the putative disease locus. A peak in the graph indicates a probable location of a disease locus, the global maximum being assumed at the maximum likelihood estimator. Figure 3.8 gives an example with a likelihood based on three marker loci. In this example the likelihood goes down steeply at two of the marker loci, because the observed segregation patterns in the data are incompatible with the causal locus being at these marker loci.

To overcome computational burden in practice one may perform multiple analyses trying to link a disease to one marker or a small group of markers at a time, rather than an overall analysis incorporating all the information. Because part of the data is common to these analyses, this leads to the problem of assigning an overall significance level to the analysis.

The likelihood inference extends to multiple disease loci, but then entails consideration of multiple recombination fractions. For linked disease loci the LOD-graph will have a multidimensional domain.

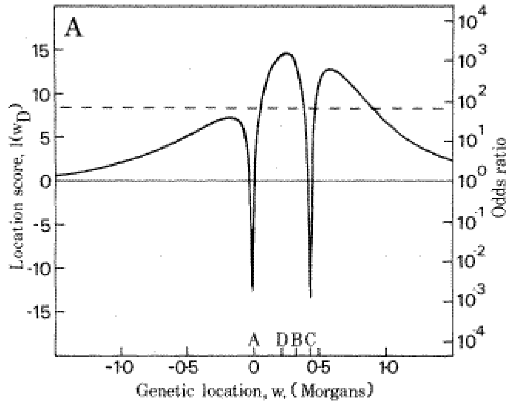

Figure 3.8. Twice the likelihood ratio (vertical scale) for parametric linkage analysis based on three marker loci, A, B, C and putative disease locus D. Data simulated to correspond to Duchenne muscular dystrophy. LOD-scores can be computed by dividing the vertical scale by $2 \log 10 \approx 4.6$. The parameter θ on the horizontal scale is defined as the recombination fraction with marker locus A transformed to genetic map coordinates using the Kosambi map function. The null hypothesis of no linkage is identified with the locus at the far left side of the horizontal axis. (Source: GM Lathrop et al. (1984). Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci.* **81**, 3443–3446.)

3.5 Counselling

Pedigree likelihoods are also the basis of *genetic counselling.* We are given a pedigree in which a phenotype of one of the members, for instance an unborn child, is unknown. We wish to assign a probability distribution to this unknown phenotype, based on all the available evidence.

The solution is simply a conditional distribution, which can be computed as the quotient of two pedigree likelihoods. The denominator is the probability of the given pedigree, annotated with all known information. The numerator is the probability of this same pedigree, but augmented with the extra information on the unknown phenotype.

3.6 Inheritance Vectors

In order to write the likelihood of an annotated pedigree, it is necessary to take into account all the possible paths by which the founder alleles are segregated through the pedigree. The "inheritance vectors" defined in this section fullfil this task. They will serve to describe Lander-Green algorithm for computing a pedigree likelihood in Section 3.8, and will later be useful for other purposes as well.

In Section 1.4 we defined a pair of *inheritance indicators* for the two meioses resulting in a zygote. Given a pedigree we can attach such a pair to every nonfounder i and locus u in the pedigree:

$$P_u^i = \begin{cases} 0, & \text{if the paternal allele of } i \text{ at locus } u \text{ is grandpaternal,} \\ 1, & \text{if the paternal allele of } i \text{ at locus } u \text{ is grandmaternal.} \end{cases}$$

$$M_u^i = \begin{cases} 0, & \text{if the maternal allele of } i \text{ at locus } u \text{ is grandpaternal,} \\ 1, & \text{if the maternal allele of } i \text{ at locus } u \text{ is grandmaternal.} \end{cases}$$

These inheritance indicators trace the two alleles of nonfounder i at a given locus back to two of the four alleles carried by his parents at that locus. Together the inheritance vectors of all nonfounders allow to reconstruct the segregation path of every allele, from founder to nonfounder. For an example see Figure 3.9, in which the founder alleles have been labelled (arbitrarily) by the numbers $1, \ldots, 8$. Every nonfounder allele is a copy of some founder allele and in the figure has received the label of the relevant founder allele. These labels can be determined by repeatedly tracing and allele back upwards from child to parent, choosing the paternal or maternal allele of the parent in accordance with the value of the inheritance indicator.

Given a pedigree with f founders and n nonfounders, and k given loci $1, \ldots, k$, we can form *inheritance vectors* by collecting the inheritance indicators of all nonfounders per locus, in the form

$$\begin{pmatrix} P_1^1 \\ M_1^1 \\ P_1^2 \\ M_1^2 \\ \vdots \\ P_1^n \\ M_1^n \end{pmatrix}, \quad \begin{pmatrix} P_2^1 \\ M_2^1 \\ P_2^2 \\ M_2^2 \\ \vdots \\ P_2^n \\ M_2^n \end{pmatrix}, \quad \ldots\ldots \quad, \begin{pmatrix} P_k^1 \\ M_k^1 \\ P_k^2 \\ M_k^2 \\ \vdots \\ P_k^n \\ M_k^n \end{pmatrix}. \tag{3.5}$$

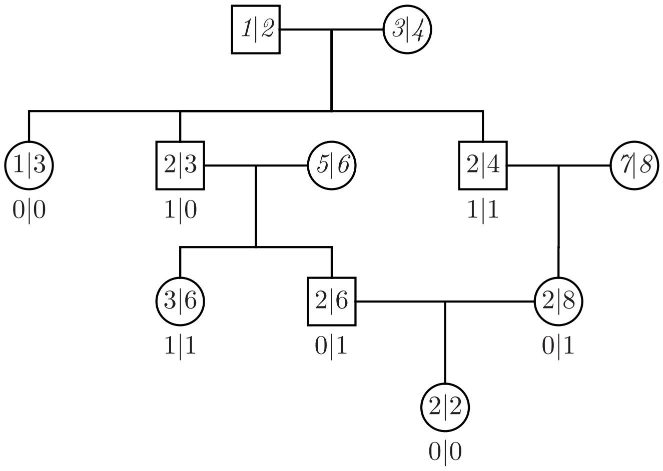

Figure 3.9. Inheritance indicators for a single locus. The founder alleles are numbered (arbitrarily) by $1, \ldots, 8$ and printed in italic inside the squares and circles. They are considered different entities, even if they may be identical in state. The nonfounder alleles are marked by the label of the founder allele and also printed inside the squares and circles. The inheritance indicators are shown below the squares and circles. Together they contribute one column vector to the inheritance matrix, with the coordinates 00101111010100.

Alternatively, we can form an *inheritance matrix* of dimension $(2n \times k)$ by considering these k vectors as the k columns of a matrix. Each row of this matrix corresponds to a (different) meiosis. As meioses are assumed stochastically independent, the rows of the matrix are independent stochastic processes.

For each locus j the f founders contribute $2f$ alleles, which are passed on to the nonfounders in the pedigree. In the absence of mutations the $2n$ alleles of the nonfounders at locus j are copies of founder alleles at locus j, typically with duplicates and/or not all founder alleles being present. The jth vector in the preceding display allows to reconstruct completely the path by which the $2f$ alleles at locus j are segregated to the nonfounders. Thus the ordered genotypes at locus j of all individuals in the pedigree are completely determined by the ordered genotypes of the founders at this locus and the jth column of the inheritance matrix.

In Section 1.4 we have seen that under the Haldane model each row of the inheritance matrix is a discrete-time Markov chain. Because the combination of independent Markov processes is again a Markov process, the vectors given in (3.5) are also a Markov chain, with state space $\{0,1\}^{2n}$. Its transition matrices can easily be obtained from the transition matrices of the coordinate processes, by the equation

$$P\left(\begin{pmatrix} P_{j+1}^1 \\ M_{j+1}^1 \\ \vdots \\ M_{j+1}^n \end{pmatrix} = \begin{pmatrix} p_{j+1}^1 \\ m_{j+1}^1 \\ \vdots \\ m_{j+1}^n \end{pmatrix} \,\Bigg|\, \begin{pmatrix} P_j^1 \\ M_j^1 \\ \vdots \\ M_j^n \end{pmatrix} = \begin{pmatrix} p_j^1 \\ m_j^1 \\ \cdots \\ m_j^n \end{pmatrix}\right)$$
$$= \prod_{i=1}^n P(P_{j+1}^i = p_{j+1}^i |\, P_j^i = p_j^i) P(M_{j+1}^i = m_{j+1}^i |\, M_j^i = m_j^i).$$

If we order the states lexicographically, then the transition matrix at locus j is the *Kronecker product* of the $2n$ transition matrices of dimension (2×2) given in (1.13). (See Section 14.13.2 for the definition of the Kronecker product. However, note that this is just naming; there is nothing to be learned beyond the preceding display.)

Typically, the inheritance process in a given pedigree is not (or not completely) observed. In the Haldane/Poisson model it is thus a *hidden Markov chain.* Observed marker or phenotypic information can be viewed as "observable outputs" of this process. There are several standard algorithms for hidden Markov processes, allowing computation of likelihoods, maximum likelihood estimation of parameters (Baum-Welch), and reconstruction of a most probable state path (Viterbi). These algorithms are described in Section 14.8, and employed by the Lander-Green algorithm, explained in Section 3.8.

Rather than considering the inheritance vectors at finitely many loci, we may think of them as processes indexed by a continuous genome. As seen in Section 1.4, the inheritance indicators $u \mapsto P_u^i$ and $u \mapsto M_u^i$ for every individual then become Markov processes in continuous time. As all meioses are independent, their combination into the vector-valued processes $u \mapsto (P_u^1, M_u^1, \ldots, P_u^n, M_u^n)$ is also a Markov process in continuous time on the state space $\{0,1\}^{2n}$.

3.7 Elston-Stewart Algorithm

The likelihood for a pedigree that is completely annotated with ordered genotypes is easy to calculate, by first multiplying the likelihoods of all founders, then going down into the pedigree and recursively multiplying the conditional likelihoods of descendants given their parents.

With F denoting the founders, NF the nonfounders and g_P^i and g_M^i the ordered genotypes of the parents of individual i, this gives an expression for the likelihood of the genotypes of the form

$$\prod_{i \in F} p(g^i) \prod_{i \in NF} p(g^i |\, g_P^i, g_M^i).$$

Here g^i is the ordered genotype of individual i, $p(g)$ is the probability that a founder has genotype g, and $p(g|\, g_P, g_M)$ is the probability that two parents with genotypes

g_P and g_M have a child with genotype g. For multilocus genotypes the latter probabilities may be sums over different inheritance patterns if one of the parents is homozygous at a locus.

This expression is next multiplied by the conditional probabilities of the phenotypes given the genotypes. Under the assumption that the individuals' phenotypes are conditionally independent given their genotypes, this results in an overall likelihood of the form

$$\prod_{i\in F\cup NF} f(x^i|\, g^i) \prod_{i\in F} p(g^i) \prod_{i\in NF} p(g^i|\, g_P^i, g_M^i).$$

Here x^i is an observed phenotype of individual i and $f(x|\, g)$ is the (penetrance) probability that an individual with genotype g has phenotype x. The independence of phenotypes given genotypes is not always realistic, but the formula can be amended for this. For instance, the assumption excludes influences from environmental factors that are common to groups of individuals.

In reality we typically observe only unordered genotypes at certain marker loci. The likelihood for the observed data is obtained by marginalizing over the unobserved data, i.e. the true likelihood is the sum of expressions as in the preceding display over all configurations of ordered genotypes and segregation paths that are compatible with the observed marker data. As we have seen in Section 3.3 the number of possible configurations may be rather large. Even for a simple pedigree as shown in Figure 3.2 and a single disease locus, there are easily 256 possibilities. To begin with we should count 2 possibilities for ordering the two alleles at each locus of each person, giving 2^{nk} possible annotated pedigrees, if there are k loci. For each locus for which the unordered genotype is not observed, the factor 2 must be replaced by l^2 for l the number of alleles for that locus. For homozygous loci there are additional possibilities hidden in the factors $p(g|\, g_P, g_M)$. It is clear that the number of possibilities increases rapidly with the number of loci and persons.

Fortunately, listing all possible pedigrees and adding their likelihoods is not the most efficient method to compute the overall likelihood for the pedigree. The *Elston-Stewart algorithm* provides an alternative method that is relatively efficient for pedigrees with many individuals and not too many loci. (For pedigrees with many loci and few individuals, there is an alternative, described in Section 3.8.) Adding the likelihoods of all possible annotated pedigrees comes down to computing the multiple sum over the genotypes of all n individuals, the overall likelihood being given by

$$\sum_{g_1}\sum_{g_2}\cdots\sum_{g_n} \prod_{i\in F\cup NF} f(x^i|\, g^i) \prod_{i\in F} p(g^i) \prod_{i\in NF} p(g^i|\, g_P^i, g_M^i).$$

Here $\sum_{g_i}$ means summing over all compatible genotypes for individual i (or summing over all genotypes with a likeihood of 0 attached to the noncompatible ones). The structure of the pedigree allows to move the sums of nonfounders to the right in this expression, computing and storing the total contributions of the individuals

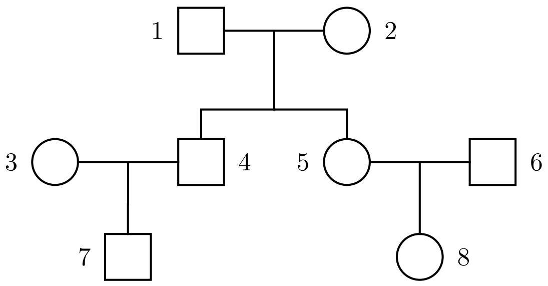

Figure 3.10. Pedigree of two nuclear families bound together used to illustrate the Elston-Stewart algorithm.

lower in the pedigree for fixed values of their ancestors, before combining them in the total.

An example makes this clearer. The pedigree in Figure 3.10 contains eight individuals. It can be viewed as consisting of the combination of the two families consisting of individuals 3, 4 and 7, and 5, 6 and 8, respectively, which are bound together by the grandparents 1 and 2. A computation of the likelihood by listing all possible pedigrees could schematically be represented by a multiple sum of the form

$$\sum_1\sum_2\cdots\sum_8\Bigl[\prod_{i=1}^8 f(i|\,i)\Bigr]\,p(1)p(2)p(3)p(6)p(4|\,12)p(5|\,12)p(7|\,34)p(8|\,56).$$

This leads to a sum of at least 2^{8k} terms (under the assumption that only the unordered genotypes are known for k loci), each of which requires 15 multiplications (not counting the multiplications and additions to evaluate the probabilities $p(i)$ and $p(j|\,k,l)$). The Elston-Stewart algorithm would rewrite this expression as

$$\begin{aligned}&\sum_1\sum_2 f(1|\,1)p(1)f(2|\,2)p(2)\times\\ &\qquad\times\Bigl[\sum_3\sum_4 f(3|\,3)p(3)f(4|\,4)p(4|\,12)\Bigl(\sum_7 f(7|\,7)p(7|\,34)\Bigr)\Bigr]\\ &\qquad\times\Bigl[\sum_5\sum_6 f(5|\,5)p(5|\,12)f(6|\,6)p(6)\Bigl(\sum_8 f(8|\,8)p(8|\,56)\Bigr)\Bigr].\end{aligned}$$

Thus the algorithm works bottom-up; it is said to be *peeling*. It first computes the contributions (between round brackets) of the individuals 7 and 8, separately, for each given value of the parents of these individuals (3 and 4 and 5 and 6, respectively). Next the algorithm moves one step higher by computing the contributions of the individuals 3 and 4, and 5 and 6, separately, for each possible value of the parents 1 and 2. Finally it combines these expressions with the contributions of the parents 1 and 2. The algorithm needs of the order $2k^2+4k^4$ additions and $2k^2+11k^4$ multiplications, well below the 2^{8k} operations if using the naive strategy.

For pedigrees that possess tree structure (no loops), the algorithm can sweep linearly through the generations and its efficiency is mainly determined by the number of loci under consideration, as these determine the size of the remaining sums. Pedigrees with some amount of inbreeding pose a greater challenge. There are many tricks by which the burden of computation can be reduced, such as factorization of the likelihood over certain loci, and efficient rules to eliminate genotypes that not consistent with the observed data. The issues here are the same as in the computation of likelihoods for *graphical models*.

3.8 Lander-Green Algorithm

The Lander-Green algorithm to compute the value of a likelihood is a particular instance of the Baum-Welch algorithm for hidden Markov models. The underlying (hidden) Markov process is the process of inheritance vectors (3.5), which determines the segregation of alleles at k loci. To ensure the Markov property we adopt the Haldane/Poisson model for the chiasmata process. For n nonfounders the states of the Markov process are vectors of length $2n$. In agreement with the general treatment of hidden Markov processes in Section 14.8 we shall denote the inheritance vectors by $Y_1,\ldots,Y_k$.

The outputs $X_1,\ldots,X_k$ of the hidden Markov model are the observed marker data of all individuals at the marker loci, and the phenotypes of all individuals at the disease locus. The marker data for a given locus (state) j are typically a vector of $n+f$ unordered pairs of alleles, one pair for every individual in the pedigree. The $2n$ alleles in this vector corresponding to nonfounders are all copies of the $2f$ founder alleles, where some founder alleles may not reappear and others appear multiple times. If the founders are chosen at random from a population that is in combined Hardy-Weinberg and linkage equilibrium, then the founder alleles are independent across loci. They are also independent of the inheritance process, as the latter depends on the meioses only. Given the inheritance process the marker data that are output at the loci are then independent. We also assume, that given the genetic information at the disease locus, the disease status (or trait value) of an individual is independent of all other variables. Under these conditions the outputs $X_1,\ldots,X_k$ fit the general structure of the hidden Markov model, as described in Section 14.8.

If the founder alleles at a marker locus j are not observed, then the output density (giving the probability of the observed markers x_j given the inheritance vector y_j at locus j) at this locus can be written as

$$q_j(x_j\,|\,y_j)=\sum P(\text{founder alleles}_j)P\bigl(x_j\,|\,y_j,\text{founder alleles}_j\bigr),$$

where the sum is over all possible ordered sets of founder alleles at locus j. Here the probabililities $P(\text{founder alleles}_j)$ can be expressed in the allele frequencies of the alleles at locus j using Hardy-Weinberg equilibrium. Furthermore, every of the probabilities $P\bigl(x_j\,|\,y_j, \text{founder alleles}_j\bigr)$ is degenerate: it is 1 if the observed marker data for locus j is compatible with the inheritance vector y_j and set of founder alleles, and 0 otherwise. This follows because the inheritance vector completely describes the segregation of the founder alleles, so that the event $\{y_j, \text{founder alleles}_j\}$ completely determines x_j. If the founder alleles are observed, then the output density is defined without the sum.

The output at the disease locus j is the vector of the phenotypes of all individuals. Its density can be written as

$$q_j(x_j\,|\,y_j) = \sum P(\text{founder alleles}_j) \prod_i f(x_j^i\,|\,\text{founder alleles}_j, y_j),$$

where x_j^i is the disease status of individual i, and f is the penetrance. This assumes that the phenotypes of the individuals are independent given their genotypes. Possible environmental interactions could be included by replacing the product by a more complicated expression.

To remain within the standard hidden Markov model set-up the outputs, including the disease phenotype, must be independent of all other variables given the state. This appears to allow diseases that depend on a single locus only?? It is not difficult, however, to extend the preceding to diseases that depend on multiple unlinked loci, e.g. on different chromosomes, which we could model with several independent Markov chains of inheritance vectors??

If the allele frequencies and penetrances are not known, then they must be (re)estimated in the M-step of the EM-algorithm, which may be computationally painful??

The underlying Markov chain consists of $2n$ independent Markov chains, each with state space $\{0,1\}$, corresponding to the $2n$ meioses in the pedigree. The chains do not have a preferred direction along the chromosome, but their initial distributions are $(\frac{1}{2},\frac{1}{2})$ and their transition probabilities are given by the recombination fractions between the loci. Thus for the $2n$-dimensional chain $Y_1,\ldots,Y_k$

$$\pi(y_1) = \Bigl(\frac{1}{2}\Bigr)^{2n},$$
$$p_j(y_j\,|\,y_{j-1}) = \theta_j^{u_j}(1-\theta_j)^{2n-u_j},$$

where $U_j = \sum_{i=1}^n |P_j^i - P_{j-1}^i| + |M_j^i - M_{j-1}^i|$ is the number of meioses that are recombinant between the loci $j-1$ and j. The recombination fraction θ_j may be known if both $j-1$ and j are marker loci. We desire to estimate it if it involves the disease locus. If the (putative) disease locus D is placed between two markers whose recombination fraction θ_{MM} is known, then we could use the relationship $\theta_{MM} = \theta_{MD}(1-\theta_{DM}) + (1-\theta_{MD})\theta_{DM}$ for the recombination fractions between the three loci to parameterize the likelihood by a single parameter. For instance, we could choose $\theta = \theta_{MD}$ as the parameter, and then have $\theta_{DM} = (\theta_{MM} - \theta)/(1-2\theta)$.

The likelihood of the hidden Markov chain contains the factors (resulting from the sequence of states marker-disease-marker)

$$\theta_{MD}^{u_{MD}}(1-\theta_{MD})^{2n-u_{MD}}\theta_{DM}^{u_{DM}}(1-\theta_{DM})^{2n-u_{DM}}.$$

In terms of the parameter θ this leads to the contribution to the log likelihood of the full data, given by

$$\begin{aligned} u_{MD}\log\theta+&(2n-u_{MD})\log(1-\theta)\\ &+u_{DM}\log\Bigl(\frac{\theta_{MM}-\theta}{1-2\theta}\Bigr)+(2n-u_{DM})\log\Bigl(1-\frac{\theta_{MM}-\theta}{1-2\theta}\Bigr). \end{aligned}$$

In the M-step of the EM-algorithm the unobserved numbers of recombinations U_{MD} and U_{DM} are replaced by their conditional expectations given the outputs, using the current estimates of allele frequencies, penetrances and recombination fractions, after which maximization over θ follows.

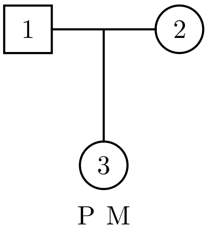

Figure 3.11. Pedigree consisting of three individuals, labelled 1, 2 ,3. The vector (P,M) is the inheritance vector of the child at a single locus.

* **3.6 Example.** As an example consider the pedigree in Figure 3.11 consisting of a father, a mother and a child (labelled 1, 2, 3), at three loci 1, 2, 3, where it is imagined that the middle locus 2 is the (putative) disease locus, which we shall assume not to be a marker locus. The corresponding hidden Markov model is pictured in Figure 3.12. The father and mother are founders and hence the states are formed by the inheritance vectors of the child, one for each locus:

$$Y_1=\begin{pmatrix}P_1\\M_1\end{pmatrix}, Y_2=\begin{pmatrix}P_2\\M_2\end{pmatrix}, Y_3=\begin{pmatrix}P_3\\M_3\end{pmatrix}.$$

The output from states 1 and 3 consists of the marker information at these loci, measured on all three individuals. The output of the disease state 2 is the phenotypic information X on all three individuals.

The hidden Markov model structure requires that the phenotype vector $X=(X^1,X^2,X^3)$ given the state Y_2 is independent of the states Y_1,Y_3 and their outputs.

To operationalize the assumption that locus 2 is the (only) disease locus (linked to loci 1 or 3) it is convenient to think of these phenotypes in the form

$$\begin{aligned} X^1 &= f(G^1_{P,2}, G^1_{M,2}, C, E^1),\\ X^2 &= f(G^2_{P,2}, G^2_{M,2}, C, E^2),\\ X^3 &= f(G^3_{P,2}, G^3_{M,2}, C, E^3). \end{aligned}$$

Here $(G^i_{P,j}, G^i_{M,j})$ is the ordered genotype of individual i at locus j, C is a "common environmental factor" that accounts for dependence and E^1, E^2, E^3 are "specific environmental factors" that account for randomness specific to the individuals. Given the state Y_j the genotype $(G^3_{P,j}, G^3_{M,j})$ of the child at locus j is completely determined by the genotypes of the parents. Consequently, given Y_2 all three phenotypes are a deterministic function of the variables $(G^1_{P,2}, G^1_{M,2}), (G^2_{P,2}, G^2_{M,2}), C, E^1, E^2, E^3$. To ensure conditional independence of these phenotypes (the output from state 2) from the other loci (states Y_1 and Y_3 and their outputs: the marker data on loci 1 and 3), we assume linkage equilibrium in the population of parents, so that $(G^1_{P,2}, G^1_{M,2}), (G^2_{P,2}, G^2_{M,2})$ are independent of the alleles $(G^1_{P,j}, G^1_{M,j}), (G^2_{P,j}, G^2_{M,j})$ for $j = 1, 3$.

With θ_1 and θ_2 the recombination fractions for the intervals 1–2 and 2–3, the likelihood can be written in the form

$$\tfrac{1}{4}\theta_1^{Z_1}(1-\theta_1)^{2-Z_1}\theta_2^{Z_2}(1-\theta_2)^{2-Z_2} q_1(O_1|\, P_1, M_1) q_2(X|\, P_2, M_2) q_3(O_3|\, P_3, M_3).$$

Here the variables $Z_j = 1_{P_j \neq P_{j-1}} + 1_{M_j \neq M_{j-1}}$ give the numbers of crossovers in the two intervals $(j = 1, 2)$, and q_j are the output densities.

The output densities for states 1 and 3 refer to the marker data O_j for the two loci and have a common form. We shall assume that the marker data on these loci consists of the unordered genotypes $\{G^1_{P,1}, G^1_{M,1}\}$ and $\{G^2_{P,1}, G^2_{M,1}\}$ of the parents and the unordered genotype $\{G^1_{P,1}, G^1_{M,1}\}$ of the child. Given the state (P_1, M_1) the latter is completely determined by the *ordered* genotypes of the parents. Therefore, the output is a sum over the probabilities of the ordered genotypes of the parents that are compatible with the observed unordered genotypes of the parents and the child. If a parent has ordered genotype (i, j) at locus 1 with probability h_{1ij}, then this yields the output density at locus 1

$$\begin{aligned} q_1(O_1|\, P_1, M_1) = \sum_{i,j}\sum_{r,s} & h_{1ij} h_{1rs} 1_{\{i,j\}=\{G^1_{P,1}, G^1_{M,1}\}} 1_{\{r,s\}=\{G^2_{P,1}, G^2_{M,1}\}} \times \\ & \times 1_{\{i(1-P_1)+jP_1, r(1-M_1)+sM_1\}=\{G^3_{P,1}, G^3_{M,1}\}}. \end{aligned}$$

Note that $i(1-P_1) + jP_1$ is simply i if $P_1 = 0$ (the grandpaternal allele) and j if $P_1 = 1$, and similarly for $r(1 - M_1) + sM_1$. Only few terms in the double sum give a nonzero contribution.

The output at locus 2 is the phenotypic vector X. If $x \mapsto f(x|\, i, j)$ is the penetrance density for an individual with ordered genotype (i, j), then the output density can be written

$$q_2(O_1|\,P_2, M_2) = \sum_{i,j}\sum_{r,s} h_{1ij}h_{1rs} f(X^1|\,i,j) f(X^2|\,r,s) \times \\ \times\, f(X^3|\,i(1-P_1)+jP_1, r(1-M_1)+sM_1).$$

This assumes that all three phenotypes are observed, and no marker data for locus 2. □

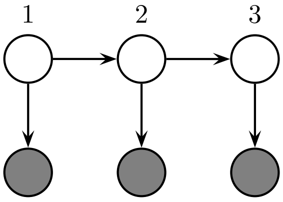

Figure 3.12. Hidden Markov model for observations on three loci.

4
Identity by Descent

Parametric linkage analysis as in Section 3.4 relies on analyzing likelihoods for observed genotypes and phenotypes for members of given pedigrees. In order to write down a likelihood it is necessary to have models for:

(i) the probabilities of the genotypes of the founders.
(ii) the segregation probabilities, describing how the founder alleles are passed on through the pedigree.
(iii) the penetrances, connecting phenotypes to genotypes.

For (i) we might assume Hardy-Weinberg and linkage equilibrium, thus describing the model completely through the allele frequencies in the population; these might be estimated from the data at hand and/or other data. For (ii) we need a map function and a model of recombination; this is the least troublesome part. The penetrances in (iii) cause no problem under an assumption of full (recessive or dominant) penetrance without phenocopies and affections that depend on a single locus, but more realistic models might require many parameters.

A full specification of the model and its parameters is referred to as *parametric linkage analysis*. In contrast, *nonparametric linkage analysis* tries to analyse pedigrees avoiding potential modelling difficulties. In particular, it tries to avoid modelling penetrances and the distribution of founder genotypes. The general idea is to sample (pedigrees of) individuals with similar phenotypes and to investigate which genes they have in common. It seems reasonable to think that these shared genes are related to the phenotype. Here “shared” is typically interpreted in the sense of “identity by descent”.

In this chapter we introduce the latter notion. We study applications to linkage of qualitative and quantatative traits in Chapters 5 and 8, respectively.

4.1 Identity by Descent and by State

Given a pedigree and a given locus, a pair of alleles of two individuals in the pedigree is called *identical by descent* (*IBD*) if they originate from (or are"physical copies" of) the same founder allele. Remember here that each founder contributes two alleles at each given locus, and all nonfounder alleles are physical copies of founder alleles, the "copying" taking place by segregation in a meiosis or a sequence of meioses.

Two alleles that are identical by descent are also *identical by state* (*IBS*) apart from mutations that may have occurred during the segregation process, meaning that they have the same genetical code. Conversely, alleles that are identical by state need certainly not be IBD. IBD-status is determined by the segregation process, not by the nature of the alleles.

Unless there is inbreeding in the pedigree, the two alleles of a single individual are never IBD, and two individuals may share 0, 1, or 2 alleles IBD, depending on chance and their family relationship. For instance, a father and child always have exactly one allele IBD, if the possibility that the father carries the maternal allele of his child is excluded. A paternal grandfather and grandchild carry 1 gene IBD if the child receives his father's paternal allele and the child's mother is not related to the grandfather.

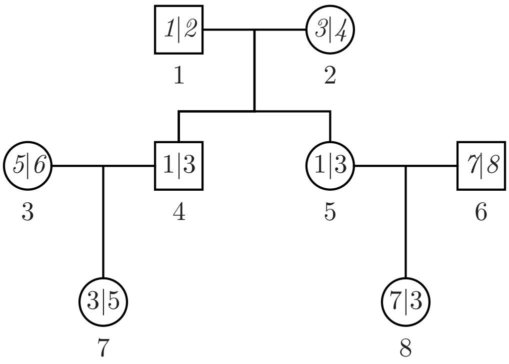

Figure 4.1. Pedigree without inbreeding. The found allels are labelled with the numbers $\mathit{1}, \mathit{2}, \ldots, \mathit{8}$ in italic. The nonfounder alleles carry the same labels in ordinary font. The vector V of alleles of the nonfounders has realization $1, 3, 1, 3, 3, 5, 7, 3$. Individuals 7 and 8 share 1 allele IBD.

The following notation is useful. For a pedigree with f founders and n nonfounders there are at each given locus $2f$ founder alleles, which "segregate" to the $2n$ nonfounder alleles. Note that the $2f$ alleles here refer to (idealized) physical entities, not to the possible genetic codes at the given locus. If we label the founder alleles arbitrarily by the numbers $1, 2, \ldots, 2f$, then we can collect full segregation information in a vector V of length $2n$, with coordinates

$$\begin{aligned} V^{2i-1} &= \text{label of paternal allele of nonfounder } i, \\ V^{2i} &= \text{label of maternal allele of nonfounder } i. \end{aligned}$$

The values of the coordinates V^i of V refer to the $2n$ nonfounder alleles; the coordinates (or indices i) themselves correspond to the nonfounder alleles. The pair of nonfounder alleles corresponding to the coordinates V^i and V^j is IBD if and only if $V^i = V^j$. Figure 4.1 shows eight founder alleles and their segregation to eight nonfounder alleles. The paternal allele of individual 7 is IBD with the maternal allele of individual 8. We shall refer to the vector V as the *segregation vector*.

An alternative notation is furnished by the inheritance vectors of the pedigree, as introduced in Section 3.6. Because the inheritance vectors completely determine the segregation of founder alleles through the pedigree, IBD-status can also be written as a function of the inheritance vectors. For large pedigrees this function is somewhat complicated, but the inheritance vectors have the advantage of having a simpler distribution. Thus we shall use the segregation vector V and inheritance vector of Section 3.6 interchangeably.

IBD-status depends on the locus. If it is important to stress this, we may label the vector V_u or the number N_u of alleles shared IBD by two individuals by a locus label u. The vectors V_{u_1} and V_{u_2}, the inheritance vectors I_{u_1} and I_{u_2}, or the numbers N_{u_1} and N_{u_2}, attached to two loci u_1 and u_2 are independent if the loci are unlinked, but they are dependent if the loci are on the same chromosome. For loci that are close together the IBD-status is the same with high probability, as it is likely that the two loci have been passed down in the pedigree without a crossover between them.

The general idea of nonparametric linkage analysis (see Chapters 5 and 8) is to find loci with relatively high IBD-values among individuals with similar phenotypes, higher than can be expected by chance. More formally, we search for loci for which the IBD-values, or equivalently the inheritance vector, is not stochastically independent of the phenotypes.

4.2 Incomplete Data

In practice, meioses are not observed and IBD-status must be inferred from the types of alleles of the individuals in the pedigree. If all founder alleles at a marker locus are different by state, then the IBD-status can be determined without error. This ideal situation is approximated by highly polymorphic markers, but in practice IBD-status is uncertain for at least some loci and individuals. The typing of additional family members may improve the situation, as this may help to resolve unknown phases. However, IBD-status involving homozygous individuals or parents can never be ascertained with certainty.

It is reasonable to replace unresolved IBD-numbers in inference by their most likely value given the observed data or, better, their conditional distributions or expectations given the observed data. IBD-status at nearby (polymorphic) loci can be almost as good, as the IBD-values at adjacent loci are the same in the absence of recombination, which is likely if the loci are close. A conditional distribution of IBD-value given the observed marker data can correctly take the probabilities of recombination into account.

Computation of such a conditional distribution requires a probability model. Because the IBD-values at a given locus are determined completely by the segregation of the founder alleles through the pedigree, the conditional distribution of IBD-status can be inferred from the conditional distribution of the inheritance vectors, introduced in Section 3.6. This is particularly attractive under the Haldane model for the chiasmata process. As explained in Section 3.6 the inheritance vectors at a sequence of ordered loci form a Markov chain $I_1, \ldots, I_k$, and the observed marker data for all individuals can be viewed as outputs $X_1, \ldots, X_k$, following the general description of a hidden Markov model in Section 14.8. Therefore, the conditional distributions of the states I_j given the outputs $X_1, \ldots, X_k$ can be computed recursively using the smoothing algorithm described in Section 14.8.

This approach does require the assumption of linkage equilibrium and uses the allele frequencies for the founders.

4.3 Distribution of IBD-indicators

The distribution of an IBD-indicator depends on the shape of the pedigree and the position of the pair of alleles therein. In this section we consider the distribution of IBD-indicators, both at a single locus and the joint distribution at multiple loci. The key is to express the IBD-values in the inheritance indicators of Section 3.6, which have a simple distribution.

4.3.1 Marginal Distributions

For a given locus u and two given individuals in the pedigree the number N_u of alleles shared IBD is a random variable that can take the values 0, 1 or 2. Its distribution does not depend on the locus u, and can of course be parametrized by two of the three probabilities $P(N_u = j)$ for $j = 0, 1, 2$. Another common parametrization is through the *kinship coefficient* and *fraternity coefficient*, defined by

$$\begin{aligned}\Theta &= \tfrac{1}{4}\mathrm{E}N_u,\\ \Delta &= P(N_u = 2).\end{aligned}$$

The kinship coefficient Θ is also the probability that a gene sampled at random from the first individual is IBD with a gene sampled at random from the second individual.

4.1 EXERCISE. Show this. [Hint: decompose the probability of this event A as $\mathrm{E}P(A|\,N_u) = 0P(N_u = 0) + \frac{1}{4}P(N_u = 1) + \frac{1}{2}P(N_u = 2)$.]

Standard family relationships, such as "sibs" or "cousins", are usually thought to have given kinship and fraternity coefficients. (See Table 4.1 for examples.) The implicit understanding here is that the individuals belong to a simple pedigree without inbreeding. The coefficients can be computed by recursive conditioning on parents, as illustrated in the following two examples.

4.2 Example (Sibs). Consider a pedigree consisting of a father, a mother and two children (two "sibs" in a "nuclear family"; see Figure 5.3). The father and mother are the founders and hence we concentrate on the IBD-status of the alleles of the children. Each child receives one allele from the father, who chooses this allele at random from his two alleles. If the father chooses the same allele for both childeren, then this paternal allele is IBD; otherwise it is not. These possibilities happen with probability $\frac{1}{2}$. Thus the variable $N_{P,u}$ defined to be 1 if the paternal allele is IBD and 0 otherwise possesses a Bernoulli distribution with parameter $\frac{1}{2}$. The same considerations are true for the maternal allele, and the corresponding variable $N_{M,u}$ is also Bernoulli distributed with parameter $\frac{1}{2}$. Furthermore, the variables $N_{P,u}$ and $N_{M,u}$ are independent.

The total number of alleles shared IBD by the two sibs is equal to $N_u = N_{P,u} + N_{M,u}$ and possesses a binomial distribution with parameters $(2, \frac{1}{2})$. It follows that the kinship and fraternity coefficients are equal to 1/4 and 1/4, respectively. □

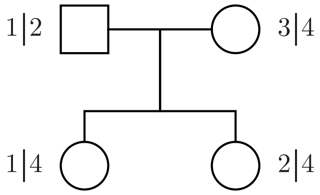

Figure 4.2. Two sibs in a nuclear family with IBD-value equal to 1 at a given locus. The alleles of the parents are labelled 1, 2, 3, 4, and the children's alleles carry the corresponding label.

4.3 Example (Cousins). Consider the two *cousins* 7 and 8 in Figure 4.1. As individuals 3 and 6 are unrelated founders, the cousins can carry at most one allele IBD: the paternal allele of cousin 7 and the maternal allele of cousin 8. Under the realization of the inheritance vectors given in Figure 4.1 these alleles are indeed IBD, but under other realizations they need not be.

It follows immediately that the fraternity coefficient Δ is equal to zero. To compute the kinship coefficient, we condition on the IBD-indicator N_u^{45} of individuals 4 and 5. The unconditional distribution of N_u^{45} is binomial with parameter $(2, 1/2)$, as in a nuclear family. Given that $N_u^{45} = 0$, the IBD-indicator N_u^{78} of individuals 7 and 8 is clearly zero; given that $N_u^{45} = 1$, the cousins have an allele in common if and only if both individual 4 and 5 segregate the allele they share IBD, with has probability 1/4; given that $N_u^{45} = 2$, the probability that the cousins have an allele in common is twice as big. Thus

$$P(N_u^{78} = 1) = \sum_{i=0}^{2} P(N_u^{78} = 1|\, N_u^{45} = i)P(N_u^{45} = i) = 0\tfrac{1}{4} + \tfrac{1}{4}\tfrac{1}{2} + \tfrac{1}{2}\tfrac{1}{4} = \tfrac{1}{4}.$$

Consequently $P(N_u^{78} = 0) = 3/4$. The kinship coefficient is $\frac{1}{4}\mathrm{E}N_u^{78} = \frac{1}{4}(3/4 * 0 + 1/4 * 1 + 0 * 2) = 1/16$. □

Relationship	Θ	Δ
Sibs	$\frac{1}{4}$	$\frac{1}{4}$
Parent-child	$\frac{1}{4}$	0
Grandparent-grandchild	$\frac{1}{8}$	0
Cousins	$\frac{1}{16}$	0
Uncle-nephew	$\frac{1}{8}$	0

Table 4.1. Kinship and fraternity coefficients of some simple pedigree relationships, under the assumption of no inbreeding.

4.3.2 Bivariate Distributions

The joint distribution of the inheritance vectors at two given loci u_1 and u_2 (two vectors as in (3.5)) can be expressed in the recombination fraction $\theta_{1,2}$ between the loci. As the IBD-values at the loci are functions of this vector, the same is true for the joint distribution of the IBD-values at two given loci, of any pair of alleles. We illustrate this by the example of two sibs.

4.4 Example (Sibs). Consider the numbers of alleles N_{u_1} and N_{u_2} shared IBD by two sibs in a nuclear family, as in Example 4.2. They are the sum of independent paternal and maternal contributions, and therefore their joint distribution can be obtained from the distribution of (N_{P,u_1}, N_{P,u_2}).

If $N_{P,u_1} = 0$, meaning that the father sends different alleles to his two children at locus u_1, then $N_{P,u_2} = 0$ if and only if both meioses are non-recombinant or if both meioses are recombinant. This, and a similar argument for the case that $N_{P,u_1} = 1$, readily shows

$$P\bigl(N_{P,u_2} = 0|\, N_{P,u_1} = 0\bigr) = (1-\theta)^2 + \theta^2,$$
$$P\bigl(N_{P,u_2} = 1|\, N_{P,u_1} = 1\bigr) = (1-\theta)^2 + \theta^2.$$

Together with the Bernoulli marginal distributions, this allows to derive the full joint distribution of N_{P,u_1} and N_{P,u_2}, as given in Table 4.2.

The vector (N_{u_1}, N_{u_2}) is distributed as the sum of two independent vectors with the distribution in Table 4.2. This leads to the joint distribution given in Table 4.3. □

$N_{P,u_1}/N_{P,u_2}$	0	1	
0	$\frac{1}{2}(1-\theta)^2 + \frac{1}{2}\theta^2$	$\theta(1-\theta)$	$\frac{1}{2}$
1	$\theta(1-\theta)$	$\frac{1}{2}(1-\theta)^2 + \frac{1}{2}\theta^2$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	1

Table 4.2. Joint distribution of IBD-counters at two loci. The parameter θ is the recombination fraction between u_1 and u_2.

N_{u_1}/N_{u_2}	0	1	2	
0	$\frac{1}{4}\psi^2$	$\frac{1}{2}\psi(1-\psi)$	$\frac{1}{4}(1-\psi)^2$	$\frac{1}{4}$
1	$\frac{1}{2}\psi(1-\psi)$	$\frac{1}{2}\bigl(1-2\psi(1-\psi)\bigr)$	$\frac{1}{2}\psi(1-\psi)$	$\frac{1}{2}$
2	$\frac{1}{4}(1-\psi)^2$	$\frac{1}{2}\psi(1-\psi)$	$\frac{1}{4}\psi^2$	$\frac{1}{4}$
	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

Table 4.3. Joint distribution of IBD-counters at two loci. The parameter ψ is defined as $\psi = (1-\theta)^2 + \theta^2$ for θ the recombination fraction between u_1 and u_2, and can be computed to be the covariance between N_{u_1} and N_{u_2}.

4.3.3 Multivariate Distributions

Inheritance indicators and IBD-indicators at multiple loci depend on the occurrence of crossovers between the loci, and therefore their joint distribution depends on the model used for the chiasmata process (cf. Theorem 1.3). In this section we restrict ourselves to the Haldane/Poisson model, which permits a simple description of the inheritance and IBD-indicators as Markov stochastic process.

In Section 1.4 it was seen that under the Haldane/Poisson model the inheritance indicators of a given meiosis indexed by continuous locus (a paternal $u \mapsto P_u$ or a maternal $u \mapsto M_u$ process) of a given meiosis are continuous time Markov processes. In a given pedigree there are multiple meioses, every one of which has an inheritance process $u \mapsto I_u^i$ attached. These processes are identically distributed and stochastically independent. The IBD-indicators of the pedigree can be expressed in the inheritance processes, and hence their distribution can be derived.

As an example consider the IBD-indicators of the alleles of two sibs in a nuclear family, as in Example 4.2. The two paternal alleles of the two sibs are IBD if and only if the two paternal meioses have the same inheritance indicator. Therefore the variable $N_{P,u}$ defined to be 1 or 0 if the paternal alleles at locus u of the sibs are IBD is distributed as the indicator of equality of two inheritance processes,

$$u \mapsto 1_{I_u^1} = I_{I_u^3}. \tag{4.5}$$

This indicator process also switches between 0 and 1, the jump times being the times where one of the two processes $u \mapsto I_u^1$ or $u \mapsto I_u^3$ switches. As both switch at the event times of a Poisson process of intensity 1 (per Morgan), the process (4.5) switches at the superposition of these two Poisson processes. By the independence of meioses the latter is a Poisson process of intensity 2 (per Morgan). In view of Lemma 1.14 the indicator process (4.5) is a Markov process with transition probabilities as in the lemma with $\lambda = 2$.

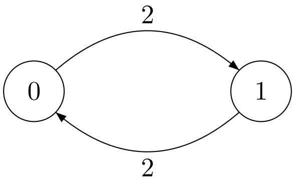

Figure 4.3. The two states and transition intensities of the Markov process $u \mapsto 1_{I_u^i = I_u^j}$, given two independent inheritance processes under the Haldane/Poisson model for crossovers.

A schematic view of the process $u \mapsto 1_{I_u^1=I_u^3}$ is given in Figure 4.3. The two circles represent the two states and the numbers on the arrows the intensities of transition between the two states. In the language of Markov processes the matrix

$$\begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix}$$

is the generator of the process (see Section 14.13).

The distributions of other IBD-indicators can be obtained similarly. For instance, the total number of alleles shared IBD by the two sibs in a nuclear family can be written as the sum of two independent processes of the preceding type, corresponding to the paternal and maternal meioses (the variables N_P and N_M of Example 4.2). For inheritance processes I^1, I^2, I^3, I^4 of four different meioses, the sum process $u \mapsto 1_{I_u^1=I_u^3} + 1_{I_u^2=I_u^4}$ has state space $\{0, 1, 2\}$ and generator matrix

$$\begin{pmatrix} -4 & 4 & 0 \\ 2 & -4 & 2 \\ 0 & 4 & -4 \end{pmatrix}.$$

Transitions from the extreme states 0 or 2 to the middle state 1 occur if one of the two indicator processes switches and hence have double intensity. A graphical representation of this Markov chain is given in Figure 4.4.

4.6 EXERCISE. Table 4.2 allows to express the conditional probability that $N_{P,u_1} = 1$ given that $N_{P,u_2} = 0$ as $2\theta(1-\theta)$. Show that under the Haldane/Poisson map function this is identical to the transition probability as in Lemma 1.14 with $\lambda = 2$.

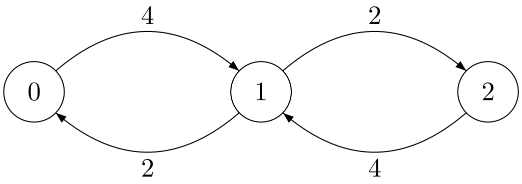

Figure 4.4. The three states and transition intensities of the Markov process $u \mapsto 1_{I_u^1 = I_u^3} + 1_{I_u^2 = I_u^4}$, given four independent inheritance processes under the Haldane/Poisson model for crossovers.

4.4 Conditional Distributions

The preceding concerns the unconditional distributions of the various processes. Nonparametric linkage analysis is based on the idea that the IBD-status at a locus and a given phenotype are stochastically dependent if and only if the locus is linked to a causal locus for the phenotype. In other words, the conditional distribution of IBD-status at a (marker) locus given the phenotype is different from the unconditional distribution if and only if the marker locus is linked to a causal locus. The size of the difference between conditional and unconditional distributions is important for the statistical power to find such a locus, and depends both on the strength of association between causal locus and phenotype, and on the distance between the marker and the causal locus.

Suppose that the phenotype depends on k causal loci in that the phenotypes of n individuals in a pedigree can be written

$$X^i = f(G^i_{\tau_1}, \ldots, G^i_{\tau_k}, C^i), \tag{4.7}$$

where the variables $G^i_{\tau_1}, \ldots, G^i_{\tau_k}$ are the genotypes at the k causal loci $\tau_1, \ldots, \tau_k$, the variables C^i account for additional random variation in the phenotypes, and f is a given function. We think of these variables as random through a combination of random sampling of founder genotypes, random meioses in the pedigree, and random "environmental influences" $C^1, \ldots, C^n$, where the three sources of randomness are assumed stochastically independent. The following theorem shows that under these assumptions the inheritance indicators at the non-causal loci depend on the phenotypes only through the inheritance indicators at the causal loci.

Let $I_u = (I_u^1, \ldots, I_u^{2n})^T$ denote the inheritance vector at locus u, which collects the inheritance indicators of the $2n$ meioses in a pedigree with n nonfounders (one of the columns in (3.5)), and for a set U of loci let $I_U = (I_u\colon u \in U)$.

4.8 Theorem. *Under (4.7) and the stated assumptions the vector $I_{\neq\tau} = (I_u\colon u \notin \{\tau_1, \ldots, \tau_k\})$ is conditionally independent of $X = (X^1, \ldots, X^n)$ given $I_\tau = (I_{\tau_1}, \ldots, I_{\tau_k})$. Consequently, for any set of loci U,*

$$P\bigl(I_U = i \,|\, X\bigr) = \sum_y P\bigl(I_U = i \,|\, I_\tau = y\bigr)\, P(I_\tau = y \,|\, X).$$

In particular, if U is a set of loci that are unlinked to the causal loci $\tau_1,\ldots,\tau_k$, then I_U is independent of $(X^1,\ldots,X^n)$.

Proof. Let G_τ^F be the genotypes of the founders of the pedigree at the loci $\tau_1,\ldots,\tau_k$ and set $C=(C^1,\ldots,C^n)$. The conditional distribution of $I_{\neq\tau}$ given X and I_τ can be decomposed as

$$P\bigl(I_{\neq\tau}=i|\,X,I_\tau\bigr)=\mathrm{E}\Bigl(P\bigl(I_{\neq\tau}=i|\,X,I_\tau,C,G_\tau^F\bigr)|\,X,I_\tau\Bigr).$$

The founder genotypes G_τ^F and the inheritance matrix I_τ completely determine the genotypes at the causal loci of all nonfounders in the pedigree. Therefore, by assumption (4.7) the vector X can be written as a function of (I_τ,C,G_τ^F), and hence can be deleted in the inner conditioning. Next we can also delete (C,G_τ^F) from the inner conditioning, as the inheritance matrices $I_{\neq\tau}$ and I_τ are completely determined by the meioses, and these are assumed independent of C and the founder genotypes. The preceding display then becomes

$$\mathrm{E}\bigl(P\bigl(I_{\neq\tau}=i|\,I_\tau\bigr)|\,X,I_\tau\bigr).$$

Here the inner probability is already a function of (X,I_τ) (and in fact of I_τ only) and hence the outer conditioning is superfluous. Thus we have proved that $P\bigl(I_{\neq\tau}=i|\,X,I_\tau\bigr)=P\bigl(I_{\neq\tau}=i|\,I_\tau\bigr)$, which implies the first assertion of the theorem (see Exercise 4.9).

The mixture representation in the second assertion is an immediate consequence of the first assertion. Furthermore, if the loci U are unlinked to the causal loci, then I_U is independent of I_τ, and the mixture representation collapses to $\sum_y P(I_U=i)\,P(I_\tau=y|\,X)=P(I_U=i)$. This shows that I_U is conditionally independent of the phenotypes X. ■

The formula given in the preceding theorem represents the conditional distribution of the inheritance matrix I_U given the phenotypes $(X^1,\ldots,X^n)$ as a mixture of the conditional distribution of I_U given I_τ, with weights the conditional distribution of I_τ given X. If U is only linked to a subset $\tau_0\subset\tau$ of the causal loci, then the unlinked causal loci can be marginalized out of the mixture and we obtain

$$P\bigl(I_U=i|\,X\bigr)=\sum_y P\bigl(I_U=i|\,I_{\tau_0}=y\bigr)\,P(I_{\tau_0}=y|\,X).$$

Thus the dependence of the inheritance process on the phenotype goes via the linked causal loci. The probabilities $P(I_U=i|\,I_{\tau_0}=y)$ of the mixed distribution are given by "transition probabilities" of the inheritance process, and do not involve the phenotypes.

Under the Haldane/Poisson model for the chiasmata the inheritance process is a Markov chain. The representation is then particularly attactive if U is linked to only a single locus τ, because the terms of the mixture are then precisely the transition probabilities of the Markov chain $u\mapsto I_u$ started at τ. If there are multiple linked causal loci, the Markov chain is conditioned to be in certain states at multiple times.

4.9 EXERCISE. Show that the following statements are equivalent ways of expressing that the random variables X and Y are conditionally independent given the random variable Z:

(i) $P(X \in A, Y \in B|\, Z) = P(X \in A|\, Z)P(Y \in B|\, Z)$ almost surely for every events A and B.

(ii) $P(X \in A|\, Y, Z) = P(X \in A|\, Z)$ almost surely for every event A.

(iii) $P(Y \in B|\, X, Z) = P(Y \in B|\, Z)$ almost surely for every event B.

5
Nonparametric Linkage Analysis

It was seen in the preceding chapter that the IBD-indicators at loci that are not linked to the causal loci are stochastically independent of the phenotype. Therefore the null hypothesis of no linkage of a given locus can be tested by testing for independence between IBD-indicator and phenotype. Nonparametric linkage analysis operationalizes this idea by comparing IBD-sharing among individuals with similar phenotypes. For an unlinked locus there should be no difference in sharing between affected and nonaffected individuals, whereas for a linked locus higher IBD-numbers among individuals with a similar phenotype are expected.

In this chapter we apply this general principle to finding loci involved in causing qualitative traits, for instance binary traits ("affected" or "nonaffected"). We consider in particular the nonparametric linkage test based on nuclear families with two affected sibs.

5.1 Nuclear Families

Consider the number N_u of alleles shared IBD at locus u by two sibs in a nuclear family, as in Figure 4.2. In Example 4.2 this variable was seen to be binomially distributed with parameters 2 and $\frac{1}{2}$

This is the correct distribution if the nuclear family is drawn at random from the population. The *affected sib pair method* is based on conditioning on the event, denoted ASP, that both children are affected. Intuitively, the information that both sibs are affected makes it more likely that they are genetically similar at the loci that are responsible for the affection. Thus the conditional distribution given ASP of the IBD-value at a locus that is linked to the disease should put more probability on the point 2 and less on the value 0. On the other hand, the conditional distribution given ASP of the IBD-value at a locus that is unrelated to the affection should be identical to the (unconditional) binomial distribution with parameters 2 and $\frac{1}{2}$.

Thus we can search for loci involved in the disease by testing whether the conditional IBD-distribution differs from the unconditional distribution.

In practice we determine the conditional IBD-distribution through random sampling from the set of nuclear families with two affected children. Let $N_u^1, N_u^2, \ldots, N_u^n$ be the numbers of alleles shared IBD at locus u by the two sibs in a random sample of n families with two affected children. These variables are a random sample from a distribution on the numbers $\{0, 1, 2\}$, given by a probability vector $z = (z_0, z_1, z_2)$ belonging to the unit simplex $S_2 = \{(z_0, z_1, z_2)\colon z_j \geq 0, \sum_j z_j = 1\}$. If the locus u is unlinked to the affection, then this distribution should not be different from the distribution found in a random sample from all nuclear families. We therefore test the null hypothesis

$$H_0\colon z = (\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}).$$

If the null hypothesis is rejected we conclude that the locus is linked to the disease. The alternative hypothesis could specify that the parameter is in the unit simplex S_2, but not equal to $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. However, under reasonable conditions it can be shown that under *ASP* the parameter z_u is always contained in the subset $H_2 = \{z \in S_2\colon 2z_0 \leq z_1 \leq \frac{1}{2}\}$, known as *Holmans' triangle*, in correspondence with our intuition that z_0 should decrease and z_2 increase under *ASP*. See Figure 5.1 and Section 5.5. Restricting the parameter set to a smaller set should make the construction of more powerful tests feasible. Moreover, even if the conditions for Holmans' triangle may not always be satisfied, it is reasonable to use a test that is powerful in particular for alternatives in the triangle.

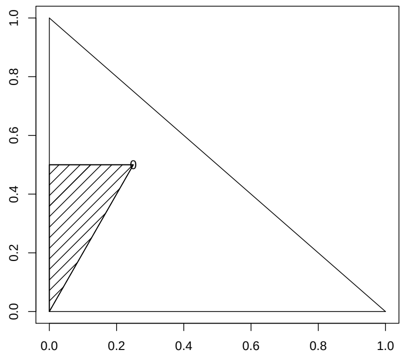

Figure 5.1. Holmans' triangle is the small, shaded triangle. Shown are the probabilities z_1 and z_2 on the horizontal and vertical axis. The null hypothesis $H_0\colon (z_1, z_2) = (1/4, 1/2)$ is the point at the upper right corner of Holmans' triangle. The large triangle is the unit simplex.

The likelihood for one family can be written as

$$z \mapsto P_z\bigl(N_u^i = j\bigr) = z_0^{1_{j=0}} z_1^{1_{j=1}} z_2^{1_{j=2}}.$$

It follows that the likelihood ratio statistic based on observing the random variables $N_u^1, N_u^2, \ldots, N_u^n$ is

$$\Lambda_u = \sup_z \frac{\prod_{i=1}^n z_0^{1_{N_u^i=0}} z_1^{1_{N_u^i=1}} z_2^{1_{N_u^i=2}}}{\prod_{i=1}^n (1/4)^{1_{N_u^i=0}} (1/2)^{1_{N_u^i=1}} (1/4)^{1_{N_u^i=2}}} = \sup_z (4z_0)^{M_{u,0}} (2z_1)^{M_{u,1}} (4z_2)^{M_{u,2}},$$

where $M_u = (M_{u,0}, M_{u,1}, M_{u,2})$ counts the number of families in which the sibs have 0, 1 or 2 alleles IBD:

$$M_{u,j} = \#\{1 \leq i \leq n : N_u^i = j\}, \qquad j = 0, 1, 2.$$

The supremum can be computed over the unit simplex S_2 or over Holmans' triangle H_2. The first possibility has the benefit of simplicity, as the maximum of the likelihood over S_2 can be seen to be taken at the point M_u/n, the observed relative frequencies of the IBD-values. Maximization over Holmans' triangle is slightly more complicated. If the unrestricted maximum likelihood estimate falls into the triangle, then the maximum value is the same as before; otherwise this must be "projected" into the triangle.

Using the full two-simplex has the further advantage that two times the log likelihood ratio statistic tends under the null hypothesis in distribution to a chisquare distribution with two degrees of freedom, as $n \to \infty$. Thus for large n a test of approximate size α is obtained by rejecting the null hypothesis if this statistic exceeds the upper α-quantile of this chisquare distribution. Because the null hypothesis $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ is at a corner of Holmans' triangle, the limit distribution of the log likelihood ratio statistic for the restricted alternative hypothesis is not chisquare, but a mixture of chisquare distributions with 0, 1, and 2 degrees of freedom. See Section 14.2. This limit distribution can still be used to determine critical values.

There are many alternatives for the likelihood ratio statistic, the most popular and simplest one being the "NPL-statistic". Under the null hypothesis the variables N_u^i are binomially distributed with parameters 2 and 1/2 and hence possess mean and variance equal to 1, whereas under the alternative their mean is bigger than 1. The variable $\sqrt{2}(N_u^i - 1)$ is therefore standardized at mean zero and variance 1 under the null hypothesis and is expected to be positive under the alternative. bigger. The *nonparametric linkage statistic* or *NPL-statistic* is defined as the scaled sum of these variables,

$$T_u = \frac{1}{\sqrt{n}} \sum_{i=1}^n \sqrt{2}(N_u^i - 1) = \sqrt{\frac{2}{n}} \bigl(M_{u,2} - M_{u,0}\bigr). \tag{5.1}$$

The null hypothesis is rejected for large values of this statistic. By the Central Limit Theorem this statistic is under the null hypothesis asymptotically standard normally distributed, so that a critical value can be obtained from the normal probability table.

5.1.1 Incomplete IBD-Information

In practice we usually do not (completely) observe the IBD-values N_u^i and hence it is necessary to extend the tests to situations with incomplete IBD-information. If X denotes the observed marker information for all individuals, then it is natural to adapt the likelihood ratio and NPL-statistics to

$$\sup_z \mathrm{E}_0\Big((4z_0)^{M_{u,0}}(2z_1)^{M_{u,1}}(4z_2)^{M_{u,2}} \,|\, X\Big) \qquad \text{and} \qquad \mathrm{E}_0(T_u|\, X).$$

The subscript 0 indicates a conditional expectation under the null hypothesis.

Assuming that $X = (X^1, \ldots, X^n)$ is a vector consisting of independent information X^i for family i and writing the likelihood ratio again as a product over the families, we can rewrite the adapted likeliood ratio statistic as

$$\begin{aligned}\sup_z \mathrm{E}_0\Big(\prod_{i=1}^n (4z_0)^{1_{N_u^i=0}}(2z_1)^{1_{N_u^i=1}}(4z_2)^{1_{N_u^i=2}} \,|\, X\Big) \\ &= \sup_z \prod_{i=1}^n \mathrm{E}_0\Big(4z_0 1_{N_u^i=0} + 2z_1 1_{N_u^i=1} + 4z_2 1_{N_u^i=2} |\, X^i\Big) \\ &= \sup_z \prod_{i=1}^n \Big(4z_0\pi_u(0|\, X^i) + 2z_1\pi_u(1|\, X^i) + 4z_2\pi_u(2|\, X^i)\Big),\end{aligned}$$

where $\pi_u(j|\, X^i) = P_0\big(N_u^i = j|\, X^i\big)$, for $j = 0, 1, 2$. We can again take the supremum over the full unit simplex or Holmans' triangle. The limit distributions are still chisquare or a mixture of chisquare distributions. See Section 14.2.

Under the same assumption on X the adapted NPL-statistic is obtained by replacing the variable N_u^i by its conditional expectation $\mathrm{E}_0\big(N_u^i|\, X_i\big) = \pi_u(1|\, X^i) + 2\pi_u(2|\, X^i)$, giving

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \sqrt{2}\big(\pi_u(1|\, X^i) + 2\pi_u(2|\, X^i) - 1\big).$$

The projection on the observed data keeps the mean value, but decreases variance, Hence it would be natural to replace the critical value by a smaller value. Because the projected statistic is still a sum of independent variables, in practice one may divide the statistic in the preceding display by the sample standard deviation of the terms of the sum and use a standard normal approximation.

The deviation from 1 of the variances of the individual terms $\sqrt{2}\big(\pi_u(1|\, X^i) + 2\pi_u(2|\, X^i) - 1\big)$ in the sum is a measure of the informativeness of the observed data X on the IBD-status at the locus u.

For computing the conditional probabilities $\pi_u(j|\, X^i)$ we can employ the hidden Markov structure discussed in Section 3.6. The IBD-values at the locus of interest can be expressed in the inheritance vectors $(P_u^1, M_u^1, P_u^2, M_u^2)^T$. Together with the inheritance vectors at the marker loci, these form the hidden Markov chain underlying the segregation process, and the observed marker data are the outputs of the chain. The smoothing algorithm for hidden Markov models yields the conditional distributions of the hidden states given the outputs.

5.2 Multiple Testing

We can apply a test for linkage of a given locus for every given locus u separately. However, in practice it is applied simultaneously for a large set of loci, typically through a plot of the graph of the test statistic against the loci (see Figure 5.2). If the graph shows a sufficiently high peak at a certain locus, then this indicates that this locus is involved in the affection.

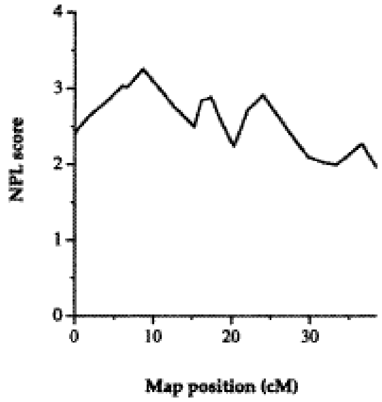

Figure 5.2. Plot of the NPL-statistic (vertical axis) versus locus (horizontal axis in Kosambi map function) for a study on schizophrenia based on 16 markers on chromosome 16p. (Source: Kruglyak at al. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Human Genetics***58**, 1347–1363.)

The question arises how high a mountain or peak should be to decide that the deviation from the null hypothesis at the given locus is significant. Writing the test statistic at locus u as T_u, we can measure this by studying the statistic

$$\sup_{u\in U} T_u,$$

where the supremum is taken over the set U of all *disease susceptibility loci* that are tested, possibly a whole chromosome. Finding a peak higher than some threshold c in the graph of $u \mapsto T_u$ is the same as this supremum statistic exceeding c. Thus a critical value c could be chosen to satisfy, for a given α,

$$P_0\Bigl(\sup_{u\in U} T_u \geq c\Bigr) \leq \alpha.$$

Of course, we have

$$\sup_{u\in U} P_0\bigl(T_u \ge c\bigr) \le P_0\Bigl(\sup_{u\in U} T_u \ge c\Bigr) \le (\#U)\; P_0\bigl(T_u \ge c\bigr).$$

The first inequality shows that the critical value should be bigger than the critical value for every given locus separately. The second inequality suggests the *Bonferroni threshold* equal to c such that $P_0\bigl(T_u \ge c\bigr) \le \alpha/\#U$ for every $u \in U$.

Unfortunately, the Bonferroni threshold is very conservative if many loci are tested. Because IBD-values at two loci are identical unless there has been a recombination between the loci, IBD-values at nearby loci are highly correlated. This typically translates into strong positive dependence between the test statistics T_u and into overlapping events $\{T_u \ge c\}$. The second inequality in the preceding display is therefore very pessimistic. Using the Bonferroni threshold would result in a *conservative test*: the level of the test will be much smaller than intended. As a result the test may easily fail to detect truly significant loci.

5.2 Example (Nuclear Families). To investigate this further consider the NPL-test statistic for nuclear families in the case of full IBD-information, given in (5.1). This was standardized to be approximately standard normally distributed under the null hypothesis for every fixed u. By similar arguments, now invoking the multivariate central limit theorem, it can be seen that the variables T_{u_i} for a given finite set of loci $u_1, \ldots, u_k$ are asymptotically jointly multivariate-normally distributed. The covariances can be computed as

$$\begin{aligned}\operatorname{cov}_0\bigl(T_{n,u_1}, T_{n,u_2}\bigr) &= 2\operatorname{cov}_0\bigl(N_{u_1}, N_{u_2}\bigr) = 4\operatorname{cov}_0\bigl(N_{P,u_1}, N_{P,u_2}\bigr)\\ &= 2\theta_{u_1,u_2}^2 + 2\bigl(1-\theta_{u_1,u_2}\bigr)^2 - 1 = 1 - 4\theta_{u_1,u_2}(1-\theta_{u_1,u_2}).\end{aligned}$$

In the second last step we have used that $\mathrm{E} N_{P,u_1} N_{P,u_2} = P(N_{P,u_1} = 1 = N_{P,u_2})$, a probability that is expressed in the recombination fraction θ_{u_1,u_2} between the loci u_1 and u_2 in Table 4.2. If the NPL-statistic is based on not too few families, we can thus act as if the stochastic process $\bigl(T_{n,u}\colon u \in U\bigr)$ is a Gaussian process with mean zero and covariance function as in the preceding display. A threshold c such that $P_0\bigl(\sup_u T_{n,u} \ge c\bigr) = \alpha$ can now be determined by simulation of the multivariate normal distribution, or (numerical) approximations to the distribution of the supremum of a Gaussian process.

Under the Haldane map function $\theta_{u_1,u_2} = \frac{1}{2}(1 - e^{-2|u_1-u_2|})$ and hence $1 - \theta_{u_1,u_2} = \frac{1}{2}(1 + e^{-2|u_1-u_2|})$. In this case the preceding display can be seen to imply

$$\operatorname{cov}_0\bigl(T_{n,u_1}, T_{n,u_2}\bigr) = e^{-4|u_1-u_2|}.$$

Together with the zero mean $\mathrm{E}_0 T_u = 0$, this shows that the limiting Gaussian process $(T_u\colon u \in U)$ is stationary. It is known as an *Ornstein-Uhlenbeck process*, and its properties are well studied in the probability literature. For instance, it can be shown that, as $c \to \infty$,

$$P\Bigl(\sup_{0\le u\le L} T_u > c\Bigr) \asymp 4L(2\pi)^{-1/2} c e^{-c^2/2}.$$

This simple approximation is accurate only for very large c, but can be improved by adding in additional terms. See Section 14.11. □

* 5.3 General Pedigrees

There are many extensions of the nonparametric linkage test to more general pedigrees than nuclear families. They all compare allele sharing at a locus among affected nonfounders in a pedigree to sharing among randomly chosen nonfounders.

Let $V_u^1, \ldots, V_u^{2n}$ be the labels of the nonfounder alleles at locus u if the founder alleles are numbered $1, 2, \ldots, 2f$. Suppose the invididuals labelled $1, 2, \ldots, n_a$ are affected. Two statistics that measure an increase in allele sharing are:

$$\sum\sum_{1\le i<j\le 2n_a} 1_{V_u^i = V_u^j},$$

$$\sum_{w_1\in\{V_u^1,V_u^2\}}\sum_{w_2\in\{V_u^3,V_u^4\}}\cdots\sum_{w_{n_a}\in\{V_u^{2n_a-1},V_u^{2n_a}\}}\prod_{j=1}^{2f}\bigl[\#(i\in\{1,\ldots,n_a\}: w_i=j)\bigr]!.$$

The first statistic is simply the total number of pairs of alleles of affected nonfounders that are IBD. The second, more complicated statistic is motivated as follows. Choose one allele from the pair of alleles of every affected nonfounder, giving labels $w_1, \ldots, w_{n_a}$; these labels are numbers in $\{1, 2, \ldots, 2f\}$, with the number j occurring $\#(i \in \{1, \ldots, n_a\}: w_i = j)$ times; compute the number of permutations of the labels $w_1, \ldots, w_{n_a}$ that keep this sequence unchanged (this is the product of factorials in the display); add these numbers of permutations over all choices of one allele. The intuition is that if many alleles are shared IBD, then the labels $w_1, \ldots, w_{n_a}$ will consist of a small number of different founder alleles, and the number of permutations leaving this vector unchanged will be large.

Given information on a random sample of pedigrees we can define an overall statistic by adding the statistics for the individual pedigrees, possibly weighted by a measure of informativeness of the pedigree. The Central Limit Theorem then implies that the statistic will be approximately normal.

To compute the distribution of test statistics of this type, it is more convenient to rewrite them as functions of the inheritance vectors $I_u = (P_u^1, M_u^1, \ldots, P_u^n, M_u^n)$ introduced in Section 3.6. Under the null hypothesis of no linkage its coordinates $P_u^1, M_u^1, \ldots, P_u^n, M_u^n$ are i.i.d. Bernoulli variables with parameter $\frac{1}{2}$. Thus the mean and variance of a test statistic of the form $T(I_u)$ can be computed as

$$\begin{aligned} \mathrm{E}_0 T(I_u) &= \sum_{i\in\{0,1\}^{2n}} \frac{1}{2^{2n}} T(i),\\ \mathrm{var}_0\, T(I_u) &= \sum_{i\in\{0,1\}^{2n}} \frac{1}{2^{2n}} T^2(i) - (\mathrm{E}_0 T)^2. \end{aligned}$$

These quantities are sufficient to obtain an approximate distribution for a (weighted) sum over pedigrees from the Central Limit Theorem. We can also obtain the limit joint distribution of the test statistic at multiple loci using

$$\mathrm{E}_0 T(I_u)\,T(I_v) = \sum_{i\in\{0,1\}^{2n}} \sum_{j\in\{0,1\}^{2n}} \frac{1}{2^{2n}} T(i)T(j) \prod_{k=1}^{2n} \theta_{u,v}^{1_{i_k\neq j_k}} (1-\theta_{u,v})^{1_{i_k=j_k}}.$$

Here $\theta_{u,v}$ is the recombination fraction between the loci u and v.

Expressing the statistics in the inheritance vectors makes it also easier to cope with incomplete IBD-information. An inheritance vector I_u of a pedigree with n nonfounders takes its values in $\{0,1\}^{2n}$, and hence a test statistic is a map $T\colon \{0,1\}^n \to \mathbb{R}$ whose values $T(i)$ are measures of compatibility of an observed value $I_u = i$ with the null hypothesis that the locus is unlinked to the affection. We rarely observe I_u itself, but must base the test on observed marker information X for the pedigree. Then it is natural to use instead the test statistic

$$\mathrm{E}_0\bigl(T(I_u)|\,X\bigr) = \sum_{i\in\{0,1\}^{2n}} T(i)\pi_u(i|\,X),$$

where $\pi_u(i|\,X) = P_0(I_u = i|\,X)$ gives the conditional distribution of the inheritance vectors given the observed data under the null distribution. Under appropriate conditions these conditional probabilities can be computed by the smoothing algorithm for hidden Markov models.

The shape of the conditional distribution $\pi_u(\cdot|\,X)$ is a measure for the informativeness of the observed marker data X. Given complete marker information this distribution is concentrated on a single point in $\{0,1\}^{2n}$, whereas a uniform distribution on $\{0,1\}^{2n}$ corresponds to no information at all. A traditional measure of "information" in a discrete distribution π on a finite set is the entropy $\sum_i \pi(i)\bigl({}^2\log \pi(i)\bigr)$. Applied in the present situation this leads to

$$\sum_{i\in\{0,1\}^{2n}} \pi_u(i|\,X)\,\bigl({}^2\log \pi_u(i|\,X)\bigr).$$

In the extreme cases of a one-point distribution or the uniform discrete distribution this reduces to 0 and $-2n$, respectively, and the number can be seen to be between these extremes for all other distributions. The measure can be used to decide whether it is useful to type additional markers in an area.

Formulation of nonparametric linkage in terms of the inheritance vectors also allows a more abstract process point of view to testing at multiple loci. Under the null hypothesis that no causal loci are linked to the loci $u_1,\ldots,u_k$ under study, the chain $I_{u_1},\ldots,I_{u_k}$ of inheritance vectors possesses a distribution that is completely determined by the stochastic model for the chiasmata. We wish to test that the observed distribution differs from this null distribution. In particular, if the chiasmata are modelled as a stationary stochastic process, then the process $I_{u_1},\ldots,I_{u_k}$ is stationary under the null hypothesis and should have a nonstationarity that is most prominent near the causal loci otherwise.

As usual this model is most attractive under the Haldane/Poisson model for the chiasmata. Under the null hypothesis the process $(I_u\colon u \in \mathbb{R})$ is then a stationary Markov chain, described in Section 4.3.3. Under the alternative the distribution, as given in Theorem 4.8, can be described as follows:

(i) The inheritance indicators at the causal loci are distributed according to some distribution $P(I_\tau = \cdot \,|\, ASP)$.

(ii) Given I_τ the remaining inheritance vectors $I_{\neq\tau} = \bigl(I_u\colon u \notin \{\tau_1, \ldots, \tau_k\}\bigr)$ are distributed according to the conditional law of $I_{\neq}$ given I_τ, independent of ASP.

This gives the image of the process $(I_u\colon u \in \mathbb{R})$ started out of stationarity at the causal loci, but evolving according to the ordinary transitions. Of course, given multiple causal loci, the fixed values of I_τ tie the process down at multiple loci and hence "evolving" has a nonlinear character.

5.4 Power of the NPL Test

The statistical power of linkage tests can be studied given a genetic model for the affection. As an example we consider the NPL for nuclear families, described in Section 5.1, under a one-locus and two-locus causal model for the disease.

The NPL-test rejects the null hypothesis that locus u is linked to the disease for large values of the statistic (5.1). This test statistic has been centered and scaled so that it has mean zero and variance 1 under the null hypothesis; for $n \to \infty$ the sequence $T_{n,u}$ tends in distribution to a standard normal distribution. The power of the test depends, at first order, on the change in the mean value under the assumption of linkage of the locus u to the disease (relative to the standard deviation).

The "changed mean" refers to the mean value $\mathrm{E}(N_u|\, ASP)$ of the IBD-indicators given the phenotype ASP. (We write N_u without the superscript i for a typical family.) These IBD-indicators can be expressed in the inheritance indicators $I_u = (P_u^1, M_u^1, P_u^2, M_u^2)^T$ of the nuclear family as

$$N_{P,u} = 1_{P_u^1 = P_u^2}, \qquad N_{M,u} = 1_{M_u^1 = M_u^2}, \qquad N_u = N_{P,u} + N_{M,u}.$$

Theorem 4.8 gives an expression for the conditional distribution of the inheritance process I_U at a set of loci U given ASP. Combining this with the preceding display we see that, for any $i, j \in \{0, 1\}$,

$$\begin{aligned} &P\bigl(N_{P,U} = i, N_{M,U} = j|\, ASP\bigr) \\ &\qquad = \sum_y P(N_{P,U} = i, N_{M,U} = j|\, I_\tau = y) P(I_\tau = y|\, ASP). \end{aligned} \tag{5.3}$$

Thus the conditional mean $\mathrm{E}(N_u|\, ASP)$ can be specified by a model for the conditional distribution of the inheritance vector I_τ at the causal loci given ASP. The conditional probabilities $P(N_{P,U}=i, N_{M,U}=j|\, I_\tau)$ can be factorized as the product $P(N_{P,U}=i|\, P_\tau^1, P_\tau^2)P(N_{M,U}=j|\, M_\tau^1, M_\tau^2)$ of the paternal and maternal meioses. For these we adopt the Haldane/Poisson model for the chiasmata process, as usual.

5.4.1 One Linked Causal Locus

In the case that U is linked to only a single causal locus τ, the preceding display can be simplified. In this situation the conditional probabilities $P(N_{P,U}=i|\, P_\tau^1, P_\tau^2)$ depend on (P_τ^1, P_τ^2) only through $N_{P,\tau}$. (The four possible configurations for (P_τ^1, P_τ^2) fall in the two groups $\{(0,0),(1,1)\}$ and $\{(0,1),(1,0)\}$ by value of $N_{P,\tau}$, and by the symmetry between the two variables P_τ^i and the symmetry in their transitions from 0 to 1 or 1 to 0 it is irrelevant for the value of $N_{P,u}$ which of the two elements of the group is the starting configuration.) This and the similar observation for the maternal indicators, allows to infer from (5.3) that

$$\begin{aligned}&P\bigl(N_{P,U}=i, N_{M,U}=j|\, ASP\bigr)\\ &=\sum_{k,l} P(N_{P,U}=i|\, N_{P,\tau}=k)P(N_{M,U}=j|\, N_{M,\tau}=l)P(N_{P,\tau}=k, N_{M,\tau}=l|\, ASP).\end{aligned}$$

The conditional distribution of the IBD-indicators given ASP can therefore be parametrized by the four probabilities $z_\tau(k,l)=P\bigl(N_{P,\tau}=k, N_{M,\tau}=l|\, ASP\bigr)$, for $k,l\in\{0,1\}$. For simplicity we assume that paternal and maternal origins are irrelevant (i.e. $z_\tau(0,1)=z_\tau(1,0)$), which leaves two degrees of freedom in the four probabilities. A convenient parameterization is given in Table 5.1. It uses the parameters δ and ε to off-set the probabilities from their null value 1/4.

Given $(N_{P,\tau}, N_{M,\tau})$ the processes $\bigl(N_{P,u}\colon u\in U\bigr)$ and $\bigl(N_{M,u}\colon u\in U\bigr)$ are independent Poisson switching processes of the type discussed in Lemma 1.14 with $\lambda=2$. In particular, the transition probability $P(N_{P,u}=j|\, N_{P,\tau}=i)$ is equal to $\psi=\frac12(1+e^{-4|u-\tau|})$ if $i=j$, and is equal to $1-\psi$ otherwise. With the parametrization given in Table 5.1 we find,

$$\begin{aligned}\mathrm{E}(N_u|\, ASP)&=2\mathrm{E}(N_{P,u}|\, ASP)=2\mathrm{E}\bigl(\mathrm{E}(N_{P,u}|\, N_{P,\tau})|\, ASP\bigr)\\ &=2(\tfrac12-\tfrac12\delta)\mathrm{E}(N_{P,u}|\, N_{P,\tau}=0)+2(\tfrac12+\tfrac12\delta)\mathrm{E}(N_{P,u}|\, N_{P,\tau}=1)\\ &=2(\tfrac12-\tfrac12\delta)(1-\psi)+2(\tfrac12+\tfrac12\delta)\psi=1+\delta(2\psi-1)=1+\delta e^{-4|u-\tau|}.\end{aligned}$$

As to be expected, the change in the mean value of the test statistic is largest at the causal locus $u=\tau$, and decreases to the null value 1 exponentially if the genetic map distance between u and the causal locus increases to infinity. By a similar argument we find that

$$\begin{aligned}\mathrm{var}(N_u|\, ASP)&=2\,\mathrm{var}(N_{P,u}|\, ASP)+2\,\mathrm{cov}(N_{P,u}, N_{M,u}|\, ASP)\\ &=2\mathrm{E}(N_{P,u}^2|\, ASP)+2\mathrm{E}(N_{P,u}N_{M,u}|\, ASP)-4\mathrm{E}(N_{P,u}|\, ASP)^2\\ &=2\bigl(\tfrac12+\tfrac12\delta(2\psi-1)\bigr)+2\bigl[(\tfrac14-\delta+\varepsilon)(1-\psi)^2\\ &\qquad+2(\tfrac14+\tfrac12\delta-\varepsilon)\psi(1-\psi)+(\tfrac14+\varepsilon)\psi^2\bigr]-4\bigl(\tfrac12+\tfrac12\delta(2\psi-1)\bigr)^2\\ &=\tfrac12-(\delta+\delta^2-2\varepsilon)(2\psi-1)^2=\tfrac12-(\delta+\delta^2-2\varepsilon)e^{-8|u-\tau|}.\end{aligned}$$

For small δ and ε this is close to the null value $1/2$.

We conclude that under the alternative the mean and variance of the NPL-statistic (5.1) are $\sqrt{2n}\delta e^{-4|u-\tau|}$ and $1-2(\delta+\delta^2-2\varepsilon)e^{-8|u-\tau|}$, respectively. If $\delta > 0$, then the test statistic tends in distribution to infinity as $n \to \infty$, whence the null hypothesis is rejected with probability tending to one. The normal approximation to the power of the test is

$$P_{\delta,\varepsilon}(T_{n,u} \geq c) \approx 1 - \Phi\Big(\frac{c - \sqrt{2n}\,\delta e^{-4|u-\tau|}}{\sqrt{1-2(\delta+\delta^2-2\varepsilon)e^{-8|u-\tau|}}}\Big).$$

For sequences of alternatives $\delta = \delta_n$ such that $\sqrt{n}\delta_n$ tends to a finite limit, this approximation tends to a limit also. For sequences of alternatives with $\sqrt{n}\delta_n \to \infty$ the normal approximation tends to 1 as does the power, but the normal approximation is not very accurate for such "large deviations" and would better be replaced by a different one.

$N_{P,\tau}/N_{M,\tau}$	0	1	
0	$\frac{1}{4} - \delta + \varepsilon$	$\frac{1}{4} + \frac{1}{2}\delta - \varepsilon$	$\frac{1}{2} - \frac{1}{2}\delta$
1	$\frac{1}{4} + \frac{1}{2}\delta - \varepsilon$	$\frac{1}{4} + \varepsilon$	$\frac{1}{2} + \frac{1}{2}\delta$
	$\frac{1}{2} - \frac{1}{2}\delta$	$\frac{1}{2} + \frac{1}{2}\delta$	

Table 5.1. Parametrization of the probabilities $z_\tau(i,j) = P(N_{P,\tau} = i, N_{M,\tau} = j|\, ASP)$. The parameter $1+\delta$ is equal to the expected value $\mathrm{E}(N_{P,\tau} + N_{M,\tau}|\, ASP)$.

The NPL-test is typically performed for multiple putative loci u simultaneously, and rejects for some locus if $\sup_u T_{n,u}$ exceeds some level. Obviously, the rejection probabilities $P_{\delta,\varepsilon}\big(\sup_u T_{n,u} \geq c\big)$ also tend to 1 as $n \to \infty$ at every fixed alternative with $\delta > 0$. To gain more insight we derive the joint limit distribution of the process $(T_{n,u}: u \in U)$ under alternatives as given in Table 5.1 with $\delta = h/\sqrt{n}$ and $\varepsilon = \varepsilon_n \to 0$. (Under the assumption that $z_\tau(0,0) \leq z_\tau(0,1)$ the convergence of ε is implied by the convergence of $\delta = \delta_n$.) By the Central Limit Theorem and the preceding calculations of mean and variance the sequence $T_{n,u}$ tends for every u to a normal distribution with mean $\sqrt{2}he^{-4|u-\tau|}$ and variance 1. The joint limit distributions follow by the multivariate Central Limit Theorem, where we need to calculate $\mathrm{cov}(N_u, N_v|\, ASP)$. It is slightly easier to compute

$$\mathrm{E}\big((N_u - N_v)^2|\, ASP\big) = 1 - e^{-4|u-v|} + (\varepsilon - 2\delta)\big(e^{-4|u-\tau|} - e^{-4|v-\tau|}\big)^2.$$

As δ and ε tend to zero this converges to $1 - e^{-4|u-\tau|}$, which is equal to $\mathrm{E}(T_u - T_v)^2$ for the Ornstein-Uhlenbeck process found in Example 5.2. We conclude that the sequence of processes $(T_{n,u}: u \in u)$ tends under the alternatives $\delta_n = h/\sqrt{n}$ and $\varepsilon_n \to 0$ in distribution to a Gaussian process $(T_u: u \in U)$ with

$$\mathrm{E}T_u = \sqrt{2}he^{-4|u-\tau|}, \qquad \text{and} \qquad \mathrm{E}T_uT_v = e^{-4|u-v|}.$$

This process is the sum of an Ornstein-Uhlenbeck process and the *drift process* $(\sqrt{2}he^{-4|u-\tau|}: u \in U)$.

5.4 EXERCISE. Verify the formula for $\mathrm{E}\bigl((N_u-N_v)^2|\,ASP\bigr)$. [Hint: It is equal to $2\mathrm{E}\bigl((N_{P,u}-N_{P,v})^2|\,ASP\bigr)+2\mathrm{E}\bigl((N_{P,u}-N_{P,v})(N_{M,u}-N_{M,v})|\,ASP\bigr)$. The first term is equal to $2\mathrm{E}(N_{P,u}-N_{P,v})^2=1-e^{-4|u-v|}$, because $|N_{P,u}-N_{P,v}|$ is independent of $N_{P,\tau}$ and hence of ASP, the conditional distribution of the process $(N_{P,u}\colon u\in U)$ given $N_{P,\tau}=0$ being the same as that of the process $(1-N_{P,u}\colon u\in U)$ given $N_{P,\tau}=1$. The second term is equal to $\mathrm{E}\bigl(\mathrm{E}(N_{P,u}-N_{P,v}|\,N_{P,\tau})\mathrm{E}(N_{M,u}-N_{M,v}|\,N_{M,\tau})|\,ASP\bigr)$, where the inner conditional expectations can be evaluated as $N_{P,\tau}(\psi_{u,\tau}-\psi_{v,\tau})+(1-N_{P,\tau})(\psi_{v,\tau}-\psi_{u,\tau})$ and the analogous expression with P replaced by M.]

5.4.2 Two Linked Loci

5.4.3 Scan Statistics

* 5.5 Holmans' Triangle

In this section we prove that the distribution of IBD-sharing of two sibs in a nuclear *ASP*-family is given by a point in Holmans' triangle, under reasonable conditions. We imagine that we sample an arbitrary nuclear family and are given the information that both sibs are affected. The (conditional) distribution of the number of alleles N_u shared IBD by the two sibs is then given by the vector of three numbers $z_u=\bigl(z_u(0),z_u(1),z_u(2)\bigr)$ defined by

$$z_u(j)=P\bigl(N_u=j|\,ASP\bigr),\qquad j=0,1,2. \tag{5.5}$$

We seek conditions under which the vector z_u is contained in Holmans' triangle.

5.6 EXERCISE. Let $h_{u,j}=P(ASP|\,N_u=j)$. Show that $2z_u(0)\le z_u(1)\le\frac12$ if and only if $h_{u,0}\le h_{u,1}\le P(ASP)$.

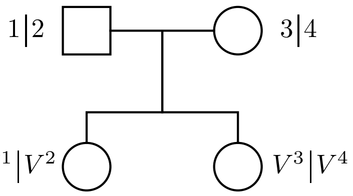

Figure 5.3. The alleles of the parents are labelled arbitrarily 1, 2, 3, 4. The children's alleles are denoted V^1,V^2,V^3,V^4, defined to be the label of the founder gene that is its origin.

We assume that the affection of the two sibs is caused by the genes at k unlinked loci. Furthermore, we assume that given their genomes the affection status of two sibs in a nuclear family is dependent only through a variable C that is independent of the genomes. More precisely, writing A^1 and A^2 for the events that sib 1 and sib 2 in a randomly chosen nuclear family are affected we assume that the conditional probability that both sibs are affected given their complete genomes $G^1 = (G^1_P, G^1_M)$, $G^2 = (G^2_P, G^2_M)$ and C can be factorized as

$$\begin{aligned} P(ASP|\, G^1, G^2, C) = P(A^1|\, G^1_{P,1}, G^1_{M,1}, \ldots, G^1_{P,k}, G^1_{M,k}, C)& \\ \times P(A^2|\, G^2_{P,1}, G^2_{M,1}, \ldots, G^2_{P,k}, G^2_{M,k}, C)&. \end{aligned}$$

The variable C may be interpreted as representing a common environment. We also assume that the penetrances $P(A^i|\, G_{P,1}, G_{M,1}, \ldots, G_{P,k}, G_{M,k}, C)$, giving the probability of affection of an individual given his genes and common environment, depend symmetrically on the gene pairs $(G_{P,j}, G_{M,j})$ at every of the k causal loci $(j = 1, 2 \ldots, k)$.

The validity of Holmans' triangle can be proved under various conditions. The simplest is to assume Hardy-Weinberg and linkage equilibrium.

5.7 Lemma. *Under the stated assumptions and combined Hardy Weinberg and linkage equilibrium the vector z_u defined in (5.5) satisfies $2z_u(0) \le z_u(1) \le \frac{1}{2}$.*

Proof. Assume first that u is one of the loci causing the affection. Define, for $i, j \in \{0, 1\}$,

$$z_u(i, j) = P\bigl(N_{P,u} = i, N_{M,u} = j|\, ASP\bigr).$$

We shall prove the inequalities

$$\begin{aligned} z_u(0,0) &\le z_u(0,1), \\ z_u(0,0) &\le z_u(1,0), \\ z_u(0,1) + z_u(1,0) &\le z_u(0,0) + z_u(1,1). \end{aligned} \tag{5.8}$$

The inequality $2z_u(0) \le z_u(1)$ then follows by taking the sum of the first two inequalities in the display, whereas the inequality $z_u(1) \le \frac{1}{2}$ follows by deducting the inequality $z_u(1) \le z_u(0) + z_u(2) = 1 - z_u(1)$ from the third.

We label the four founder alleles (arbitrarily) by $1, 2, 3, 4$, and define the vector $V_u = (V_u^1, V_u^2, V_u^3, V_u^4)^T$ as the labels of the two alleles at locus u of the first sib (V_u^1, V_u^2) and of the second sib (V_u^3, V_u^4), respectively. The IBD-indicators are $N_{P,u} = 1_{V_u^1 = V_u^3}$ and $N_{M,u} = 1_{V_u^2 = V_u^4}$. There are 16 possible configurations for the vector V_u, listed in Table 5.2 together with the values of the induced IBD-indicators. The values fall in four groups of four, which we denote by $C_{0,0}, C_{0,1}, C_{1,0}$ and $C_{1,1}$.

Let $\tau_1, \ldots, \tau_k$ be the causal loci and let V be the $(4 \times k)$-matrix with columns $V_{\tau_1}, \ldots, V_{\tau_k}$,. Furthermore, let G the $(4 \times k)$-matrix with columns the the four alleles of the father and mother at locus τ_j. As always these matrices are independent, because segregation is independent of founder genotypes, and the rows of V are independent as the four meioses between the two parents and their children are assumed to be independent. In the present situation, all elements of V are stochastically independent in view of the additional assumption that the k causal loci are unlinked, and all elements of G are independent by the equilibrium assumptions. By Bayes' formula, for any v,

$$P(V=v|\,ASP) = \frac{P(ASP|\,V=v)P(V=v)}{P(ASP)}.$$

The probabilities $z_u(i,j)$ are the sums of the left side over the sets of v such that the column v_u in v corresponding to locus u (assumed to be one of the τ_j) belongs to $C_{i,j}$.

The probability $P(V=v)$ is equal to $(2^{-4})^k$, independent of v, because the causal loci are assumed to be unlinked. To establish the inequalities in (5.8) if suffices to compare expressions of the type

$$\sum_{v:v_u\in C} P\bigl(ASP|\,V=v\bigr).$$

For instance, we prove the first inequality by showing that this expression is smaller for $C=C_{0,0}$ than for $C=C_{0,1}$, and for the third inequality we compare the expression for $C=C_{0,1}\cup C_{1,0}$ and $C=C_{0,0}\cup C_{1,1}$.

The vector V completely describes how the founder alleles G segregate to the two sibs, and hence given G the event $V=v$ completely determines the genes of the two sibs. In fact, given $V=v$ the first sib has alleles $G_{v_{1j},j}, G_{v_{2j},j}$ at locus τ_j, and the second sib $G_{v_{3j},j}, G_{v_{4j},j}$. Combining this with the assumption, we can write, for every fixed v,

$$\begin{aligned} P(ASP|\,V=v,G,C) &= P\bigl(A^1|\,G_{v_{11},1},G_{v_{21},1},\ldots,G_{v_{1k},k},G_{v_{2k},k},C\bigr)\\ &\qquad\times P\bigl(A^2|\,G_{v_{31},1},G_{v_{41},1},\ldots,G_{v_{1k},k},G_{v_{3k},k},C\bigr).\end{aligned}$$

The expected value of the right side with respect to (G,C) is the probability $P\bigl(ASP|\,V=v\bigr)$, for every fixed v.

For simplicity of notation assume that the locus u of interest is the first locus τ_1, and denote the second to last columns $(v_{ij})_{j>1}$ of v by w. For given w abbreviate

$$\begin{aligned} f_{ij} &= P\bigl(A^1|\,G_{i1},G_{j1},G_{v_{12},2},G_{v_{22},2}\ldots,G_{v_{1k},k},G_{v_{2k},k},C\bigr),\\ g_{ij} &= P\bigl(A^2|\,G_{i1},G_{j1},G_{v_{32},2},G_{v_{42},2}\ldots,G_{v_{3k},k},G_{v_{4k},k},C\bigr).\end{aligned}$$

Then we find

$$\begin{aligned} \sum_{v:v_u\in C_{0,0}} P\bigl(ASP|\,V=v\bigr) &= \sum_w \mathrm{E}(f_{13}g_{24}+f_{23}g_{14}+f_{14}g_{23}+f_{24}g_{13}),\\ \sum_{v:v_u\in C_{0,1}} P\bigl(ASP|\,V=v\bigr) &= \sum_w \mathrm{E}(f_{13}g_{23}+f_{14}g_{24}+f_{23}g_{13}+f_{24}g_{14}),\\ \sum_{v:v_u\in C_{1,0}} P\bigl(ASP|\,V=v\bigr) &= \sum_w \mathrm{E}(f_{13}g_{14}+f_{14}g_{13}+f_{23}g_{24}+f_{24}g_{23}),\\ \sum_{v:v_u\in C_{1,1}} P\bigl(ASP|\,V=v\bigr) &= \sum_w \mathrm{E}(f_{13}g_{13}+f_{23}g_{23}+f_{14}g_{14}+f_{24}g_{24}),\end{aligned}$$

where the expectation is over the (hidden) variables G and C, and the sums on the right side are over the (hidden) indices v_{ij} with $j > 1$. The inequalities in the lemma are comparisons of these four sums.

The third sum minus the first sum is proportional to $z_u(1,0) - z_u(0,0)$ and can be written in the form

$$(5.9) \qquad \sum_w \mathrm{E}(f_{13} - f_{23})(g_{14} - g_{24}) + \sum_w \mathrm{E}(f_{14} - f_{24})(g_{13} - g_{23}).$$

Both terms in this sum are nonnegative, as can be seen as follows. in case of the first sum. The variables $G_{3,1}$ and $G_{4,1}$ occur in the first and second term of the product $(f_{13} - f_{23})(g_{14} - g_{24})$, respectively, and not in the other term. Because all variables G_{ij} are independent we may compute the expectations relative to these variables first, for fixed values of the variables $G_{1,1}$ and $G_{2,1}$ and the (hidden) variables $G_{i,j}$ with $j > 1$. If there is only one locus ($k = 1$), then $f_{ij} = g_{ij}$ and computing the expectation relative to G_{31} and G_{41} collapses the expression to the expectation of a square, which is nonnegative. If there is more than one locus involved, then the conditional expectations

$$\mathrm{E}\bigl(f_{13} - f_{23} |\, G_{1,1}, G_{2,1}, G_{i,j}, j > 1\bigr)$$
$$\mathrm{E}\bigl(g_{14} - g_{24} |\, G_{1,1}, G_{2,1}, G_{i,j}, j > 1\bigr)$$

are different, as the functions f_{ij} and g_{ij} may depend on different variables at the loci $\tau_2, \ldots, \tau_k$. However, we can perform the same reduction by integrating out variables that occur in only one term of the product. The resulting expression is a square, and hence has nonnegative expectation.

The proof of the first inequality in (5.8) is the same, after permuting indices. To prove the third inequality we subtract the sum of the second and third sums from the sum of the first and fourth sums to obtain

$$(5.10) \qquad \sum_w \mathrm{E}(f_{13} + f_{24} - f_{14} - f_{23})(g_{13} + g_{24} - g_{14} - g_{23}).$$

Reasoning as before this can be reduced to the expectation of a square, which is nonnegative.

This concludes the proof if u is one of the disease loci. For a general locus u we decompose the probabilities of interest relative to the disease loci $\tau_1, \ldots, \tau_k$ as

$$\begin{aligned} z_u(j) = \sum_{j_1} \cdots \sum_{j_k} & P\bigl(N_u = j |\, N_{\tau_1} = j_1, \ldots, N_{\tau_k} = j_k, ASP\bigr) \\ & \times P\bigl(N_{\tau_1} = j_1, \ldots, N_{\tau_k} = j_k |\, ASP\bigr). \end{aligned}$$

Here the event ASP in the first conditional probability on the right can be removed, as the phenotypic information ASP is not informative on the segregation at locus u given the segregation information about the disease locations $\tau_1, \ldots, \tau_k$. (Indeed, given the IBD-indicators at the disease loci the IBD-indicators at the locus of interest u are determined by crossovers between u and the disease loci. The genotypes at the disease loci are not informative about the crossover process.) Also because the disease loci are unlinked, the locus u can be linked to at most one of the k disease loci $\tau_1, \ldots, \tau_k$, so that all except one IBD-indicator, say N_{τ_1}, can be removed in the first probability. We can next sum out the IBD-indicators at $\tau_2, \ldots, \tau_k$, obtaining the equation

$$\begin{pmatrix} z_u(0) \\ z_u(1) \\ z_u(2) \end{pmatrix} = A \begin{pmatrix} z_{\tau_1}(0) \\ z_{\tau_1}(1) \\ z_{\tau_1}(2) \end{pmatrix},$$

for A the (3×3)-matrix of probabilities $A_{j,j_1} = P\bigl(N_u = j |\, N_{\tau_1} = j_1\bigr)$. This matrix can be derived from Table 4.3, where $\psi = \theta^2 + (1-\theta)^2$. By the first part of the proof the vector on the far right (or rather the vector of its first two coordinates) is contained in Holmans' triangle. It can thus be written as the convex combination of the three extreme points of the triangle. Thus we can write the right side in the form, for some element $(\lambda_0, \lambda_1, \lambda_2)$ of the unit simplex in $\mathbb{R}^3$,

$$\begin{aligned} &A\left(\lambda_0 \begin{pmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} + \lambda_1 \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \\ \frac{1}{4} \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right) \\ &\qquad = \lambda_0 A \begin{pmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} + \lambda_1 A \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \\ \frac{1}{4} \end{pmatrix} + \lambda_2 A \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \end{aligned}$$

From the fact that $\frac{1}{2} \leq \psi \leq 1$ we can infer that the three vectors on the far right are contained in Holmans' triangle. The same is then true for their convex hull. ■

Inspection of the preceding proof shows that Hardy-Weinberg and linkage equilibrium are used to ensure that the expectations (5.9)-(5.10) are nonnegative. Nonnegativity is also reasonable without equilibrium assumptions. For instance, in the case of an affection caused by a single locus it suffices that there exists an ordering of the alleles such that the map

$$g \mapsto P(A|\, G_P = g, G_M, C) \tag{5.11}$$

is nondecreasing. In order words, the alleles can be ordered in their severity for the affection.

In the case of multiple loci similar, but more complicated assumptions can be made.

V^T	N_P	N_M	N
1324	0	0	0
2314	0	0	0
1423	0	0	0
2413	0	0	0
1323	0	1	1
1424	0	1	1
2313	0	1	1
2414	0	1	1
1314	1	0	1
1413	1	0	1
2324	1	0	1
2423	1	0	1
1313	1	1	2
2323	1	1	2
1414	1	1	2
2424	1	1	2

Table 5.2. Possible values of the inheritance vector V at a single locus for a nuclear family together with the induced IBD-values.

6
Genetic Variation

In this chapter we study decomposition of (co)variances of *quantitative traits*, phenotypes that are quantitative variables. This is of interest in its own respect, and is also the basis for regression models that explain covariation of quantitative traits through genetic factors. The latter models are used in Chapter 8 to discover quantitative trait loci.

Most quantitative traits with a genetic component depend also on environmental influences. Genetic and environmental effects are often modelled as independent random variables, and often it is also assumed that the two influences contribute additively. A quantitative trait X can then be written as a function

$$X = f(G) + E,$$

of genes G and environment E. For a randomly chosen person from the population the variables G and E are usually assumed independent.

In this chapter we focus on the dependence of a trait on genetic factors only, and hence consider functions $X = f(G)$. We consider the distribution of the trait X for a person drawn at random from a population, or the joint distribution for a pair of persons belonging to a given pedigree.

Throughout the chapter we assume that the population is in combined Hardy-Weinberg and linkage equilibrium.

6.1 Variance

Consider a trait X that depends on the genes at k loci in the form

$$X = f(G_{P,1}, \ldots, G_{P,k}, G_{M,1}, \ldots, G_{M,k}) = f(G_P, G_M).$$

Here $(G_{P,i}, G_{M,i})$ is the ordered genotype at the ith locus, and $G_P = G_{P,1} \cdots G_{P,k}$ and $G_M = G_{M,1} \cdots G_{M,k}$ are the paternal and maternal haplotypes. For simplicity, we assume that the paternal or maternal origin of the alleles is irrelevant and that the unordered gene pairs influence the trait rather than haplotypes. This means that the function f is invariant under exchanging its ith and $(i+k)$th arguments.

As there are only finitely many possible values for the alleles, there are only finitely many possible trait values. Because the number of values can be large, it will still be useful to replace X by suitable approximations. The simplest approximation is the population mean. This is the mean value $\mathrm{E}X$ of the trait X of a randomly chosen individual from the population, and is the best constant approximation in the sense of minimizing the square expectation $\mathrm{E}(X-c)^2$ over all constants c. A reasonable approximation that is both simple and does use the genotypes is an additive variable of the form $\sum_i f_i(G_{P,i}) + \sum_i f_i(G_{M,i})$. It is reasonable to determine suitable functions f_i also by minimizing the square expectation of the residual. This *additive approximation* does not allow "interactions" between the alleles. This can be remedied also considering sums of functions of two alleles, or three alleles, etc., yielding a sequence of increasingly accurate, but also more complicated, approximations.

As we assume that the population is in equilibrium and the trait X refers to a random person in the population, the alleles $G_{P,1}, G_{M,1}, G_{P,2}, G_{M,2}, \ldots, G_{P,k}, G_{M,k}$ are independent random variables. We can therefore apply the standard Hoeffding decomposition (see Section 14.6) to X viewed as a function of these $2k$ independent random variables to compute the various terms of the approximations. (The equilibrium (or independence) assumption is handy to simplify formulas, but the program could in principle also be carried out under other assumptions on the joint distirbution of the allleles.) Because the $2k$ variables form k natural pairs, it is useful to group the various terms in the decomposition both by order and by reference to the pairs.

The linear part of the decomposition can be grouped by locus, and takes the form

$$\sum_{i=1}^{k} \bigl(f_i(G_{P,i}) + f_i(G_{M,i})\bigr),$$

for the functions f_i given by

$$f_i(g) = \mathrm{E}(X|\, G_{P,i} = g) - \mathrm{E}X. \tag{6.1}$$

The same function f_i is applied to the paternal and the maternal alleles $G_{P,i}$ and $G_{M,i}$, in view of the assumed symmetry between these variables (the distributions of $(X, G_{P,i}, G_{M,i})$ and $(X, G_{M,i}, G_{P,i})$ are the same, and $f(G_P, G_M)$ remains the same upon exhcanging $G_{P,i}$ and $G_{M,i}$). The value $f_i(g)$ is known as the *breeding value* of allele g at locus i, or also the *average excess* of the allele. The sum of the breeding values over the loci is the overall breeding value.

The pairs in the quadratic part of the Hoeffding decomposition can be grouped according to whether they refer to the same locus or to different loci. The quadratic part can be written as

$$\sum_{i=1}^{k} f_{ii}(G_{P,i}, G_{M,i}) + \sum\sum_{1\le i<j\le k} \bigl(f_{ij}(G_{P,i}, G_{P,j}) + f_{ij}(G_{M,i}, G_{M,j}) \\ + f_{ij}(G_{P,i}, G_{M,j}) + f_{ij}(G_{M,i}, G_{P,j})\bigr),$$

for the functions f_{ii} and f_{ij} (with $i \neq j$) given by

$$\begin{aligned} f_{ii}(g,h) &= \mathrm{E}(X|\, G_{P,i} = g, G_{M,i} = h) \\ &\qquad - \mathrm{E}(X|\, G_{P,i} = x) - \mathrm{E}(X|\, G_{M,i} = h) + \mathrm{E}X, \end{aligned} \tag{6.2}$$

$$\begin{aligned} f_{ij}(x,y) &= \mathrm{E}(X|\, G_{P,i} = g, G_{M,j} = h) \\ &\qquad - \mathrm{E}(X|\, G_{P,i} = g) - \mathrm{E}(X|\, G_{M,j} = h) + \mathrm{E}X. \end{aligned} \tag{6.3}$$

The terms of the form $f_{ii}(G_{P,i}, G_{M,i})$ correspond to interactions within a locus, and are called *dominance* interactions. The other terms correspond to interactions between loci, and are referred to as *epistasis*. By the assumed symmetry and equalities in distribution, the function in (6.3) remains the same if P and/or M is replaced by M and/or P. Thus only a single function $f_{i,j}$ arises for every pair (i,j).

The higher order interactions can also be partitioned in groups. Third order interactions could refer to three different loci, or two loci, whereas fourth order interactions could refer to four, three or two loci, etc. It is not particularly interesting to give names to all of these, as they are usually neglected.

The variance of the trait can be decomposed as

$$\operatorname{var} X = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \cdots,$$

where σ_A^2 is the variance of the linear term, σ_D^2 the variance of the sum of the dominance interactions and σ_{AA}^2 the variance of the sum of epistatic interactions. These appear to be usual notations, together with notations such as σ_{AAD}^2, σ_{DD}^2, σ_{AAA}^2, etc. for higher order interactions. Note that σ_D^2 and σ_{AA}^2 both refer to pairwise interactions, even though their numbers of subscripts might suggest differently: each subscript "A" refers to a single allele at a (different) locus, whereas a symbol "D" refers to the pair of alleles at a locus.

The Hoeffding decompositions are a sequence of progressively more complicated approximations to the phenotype X, giving best approximations in terms of square expectation (see Section 14.6). In the present setting the square expectation refers to the randomness inherent in sampling an individual from the population and hence can be viewed as a "population mean square error". Consequently, the terms of the decomposition depend on the population characteristics, such as the frequencies of the various alleles in the population. This is clear for the zero-order approximation, which is just the population mean of the phenotype, but it is sometimes forgotten for the more complicated higher-order approximations. In particular, the first-order approximation gives an additive model

$$\mathrm{E}X + \sum_{i=1}^{k} \bigl(f_i(G_{P,i}) + f_i(G_{M,i})\bigr)$$

for the phenotype. This is easily misinterpreted as a *causal* linear model, in the sense that if one would create an individual from alleles $G_{P,1}, G_{M,1}, \ldots, G_{P,k}, G_{M,k}$, then the model would correctly give her phenotype. Although the formula may give some indication of this phenotype, this is not a correct usage of the model. The Hoeffding decomposition yields an additive model that is optimal (among additive models) on the average in the population. Had the individual been chosen at random from the given population, then substituting his alleles in the first-order approximation would make some sense. Better, had *many* individuals been chosen at random from the population and their phenotypes been predicted by substituting their phenotypes in the formula, then this would have made good sense "on the average". For any particular individual the formula does not necessarily give a sensible outcome. This is further illustrated for a monogenetic trait in the next section.

6.1.1 Monogenetic, Biallelic Traits

For a monogenic trait, given through a function $X = f(G_P, G_M)$ of the ordered genotype (G_P, G_M) at a single locus, the only interaction is the dominance interaction. The Hoeffding decomposition takes the form

$$X = \mathrm{E}X + f_1(G_P) + f_1(G_M) + f_{11}(G_P, G_M).$$

The corresponding variance decomposition is $\operatorname{var} X = \sigma_A^2 + \sigma_D^2$, for $\sigma_A^2 = 2\mathrm{E}f_1^2(G_P)$ the additive variance and $\sigma_D^2 = \mathrm{E}f_{11}^2(G_P, G_M)$ the dominance variance.

In the special case of a biallelic gene, with alleles A_1 and A_2, the trait X can have only three different values: $f(A_1, A_1)$, $f(A_1, A_2) = f(A_2, A_1)$ and $f(A_2, A_2)$. We can define the *effect* of allele A_2 over allele A_1 as $a = \frac{1}{2}\bigl(f(A_2, A_2) - f(A_1, A_1)\bigr)$, and introduce a second parameter k so that

$$\begin{aligned} f(A_1, A_1) &= f(A_1, A_1), \\ f(A_1, A_2) = f(A_2, A_1) &= f(A_1, A_1) + (1 + k)a, \\ f(A_2, A_2) &= f(A_1, A_1) + 2a. \end{aligned} \tag{6.4}$$

We consider the model as strictly *additive* if $k = 0$, because in this case each allele A_2 adds an amount a to the trait value relative to the base value for allele A_1. The values $k = -1$ and $k = 1$ correspond to the allele A_2 being *recessive* or *dominant*, respectively, as the trait of a heterozygous individual (A_1, A_2) is equal to that of the monozygous individual (A_1, A_1) or (A_2, A_2) in that case. In general the parameter k is not restricted to $\{-1, 0, 1\}$ and may assume noninteger values, with values strictly less than -1 (*underdominance*) and strictly bigger than 1 (*overdominance*) being not excluded.

The parameters a and k are called the *homozygous effect* and the *dominance coefficient*, respectively, and permit a useful reparametrization of the relative positions of the three values $f(A_1, A_1)$, $f(A_1, A_2) = f(A_2, A_1)$ and $f(A_2, A_2)$. (We need a third parameter to describe the absolute positions, but this is irrelevant if we are interested only in variances.) The Hoeffding decomposition can be expressed in these parameters and the allele frequencies. If A_1 and A_2 have population frequencies p_1 and $p_2 = 1 - p_1$, respectively, then, under Hardy-Weinberg equilibrium the frequencies of the unordered genotypes A_1A_1, A_1A_2 and A_2A_2 are $p_1^2, 2p_1p_2$ and p_2^2, and hence

$$\begin{aligned}
\mathrm{E}X &= f(A_1,A_1) + 2p_1p_2(1+k)a + p_2^2 2a,\\
\mathrm{E}(X|\,G_P = A_1) &= f(A_1,A_1) + p_2(1+k)a,\\
\mathrm{E}(X|\,G_P = A_2) &= f(A_1,A_1) + p_1(1+k)a + p_2 2a.
\end{aligned}$$

From this it follows by straightforward algebra (using definitions (6.1)–(6.3) summarized in Table 6.1) that

$$\begin{aligned}
f_1(A_1) &= -ap_2\bigl(1+(p_1-p_2)k\bigr),\\
f_1(A_2) &= ap_1\bigl(1+(p_1-p_2)k\bigr),\\
f_{11}(A_1,A_1) &= -2ap_2^2k,\\
f_{11}(A_1,A_2) = f_{1,1}(A_1,A_2) &= 2ap_1p_2k,\\
f_{11}(A_2,A_2) &= -2ap_1^2k,
\end{aligned}$$

The additive variance and dominance variance can be expressed in the new parameters as

$$\begin{aligned}
\sigma_A^2 &= 2\bigl(p_1f_1(A_1)^2 + p_2f_1(A_2)^2\bigr) = 2a^2p_1p_2\bigl(1+(p_1-p_2)k\bigr)^2,\\
\sigma_D^2 &= (2p_1p_2ak)^2.
\end{aligned}$$

The dominance variance σ_D^2 is zero if the genes act in a strictly additive fashion ($k=0$). Conversely, zero dominance variance implies additivity unless the other parameters take extreme (uninteresting) values ($p_1 \in \{0,1\}$ or $a=0$). To judge the relative contributions of additive and dominance terms if $k \neq 0$, we can compute the quotient of the additive and dominance variances as

$$\frac{\sigma_A^2}{\sigma_D^2} = \frac{\bigl(1+(p_1-p_2)k\bigr)^2}{2p_1p_2k^2}.$$

This clearly tends to infinity if $k \to 0$, but is also very large if one of the alleles is rare ($p_1 \approx 0$ or 1). Thus the relative contributions of additive and dominance terms is a somewhat complicated function of both the dominance effect and the allele frequencies.

We conclude that two possible definitions of *dominance*, through the magnitude of the dominance variance σ_D^2 or through the deviation of the parameter k from zero, do not always agree. The parameter k possesses a causal interpretation, as it directly links phenotype to genotype through the formulas (6.4). The apparent contradiction arise, because the dominance variance σ_D^2 is relative to the population.

To explain this more clearly we express the additive approximation to X in the number of A_2-alleles $S = 1_{G_P=A_2} + 1_{G_M=A_2}$ as

$$\begin{aligned}
\mathrm{E}X + f_1(G_P) + f_1(G_M) &= \mathrm{E}X + (2-S)f_1(A_1) + Sf_1(A_2)\\
&= \mathrm{E}X + 2f_1(A_1) + a\bigl(1+(p_1-p_2)k\bigr)S.
\end{aligned}$$

G_P	G_M	freq	X	$\mathrm{E}(X - \mathrm{E}X \mid G_P)$	$\mathrm{E}(X - \mathrm{E}X \mid G_M)$
A_1	A_1	p_1^2	$f(A_1, A_1)$	$f_1(A_1)$	$f_1(A_1)$
A_1	A_2	p_1p_2	$f(A_1, A_2)$	$f_1(A_1)$	$f_1(A_2)$
A_2	A_1	p_2p_1	$f(A_2, A_1)$	$f_1(A_2)$	$f_1(A_1)$
A_2	A_2	p_2^2	$f(A_2, A_2)$	$f_1(A_2)$	$f_1(A_2)$

Table 6.1. Values of a monogenetic trait $X = f(G_P, G_M)$ depending on a biallelic gene with alleles A_1 and A_2 with frequencies p_1 and p_2 in the population.

The linear function of S on the right is actually the function $\alpha + \beta S$ that minimizes $\mathrm{E}(S - \alpha - \beta S)^2$ over all linear functions (the "linear regression of X on S.) Besides by direct calculation, we can see this from the fact that, conversely, any linear function of S can be written in the form $g(G_P) + g(G_M)$ for some function g (see Problem 6.5), so that the right side of the display inherits the property of being a best least squares approximation of X from the variable $f_1(G_P) + f_1(G_M)$.

From a causal point of view it is counter-intuitive that for fixed values of a and k, the slope $a\bigl(1+(p_1-p_2)k\bigr)$ of this linear regression can vary, and can even be both positive and negative, depending on the allele frequencies. This is explained by the fact that the linear approximation is population based. If one of the alleles is rare, then it receives little weight in the regression. Reversion of the slope of the regression line may then occur in case of underdominance ($k < -1$) or overdominance ($k > 1$).

Figure 6.1 illustrates this observaton. In each pair of a left and a right panel the causal parameter k is the same, but the allele frequencies differ, with p_2 equal to 0.5 in the left panels and equal to 0.9 in the right panels. The causal parameter varies from top to bottom, taking the values $k = 0, 0.5$, and 2. The horizontal axis shows the value of S, which can assume only the values $0, 1, 2$, but is visualized as a continuous variable. The three different phenotypes (for $S = 0, 1, 2$) are visualized by the vertical heights of three asterisks. In the two top panels the causal effect is exactly additive ($k = 0$), and the two regression lines are the same and follow the asterisks. In the two middle panels the causal effect is increasing as S increases from 0 to 1, but superadditive, with as a result that the least square fits are not the same. The third panel shows a causal interaction between the alleles: the phenotype of heterozygotes ($S = 1$) is higher than the phenotypes of both types of homozygotes ($S = 0$ and $S = 2$). The regression lines can of course not reflect this interaction, straight as they are. Remarkably, the slopes of the regression lines in the left and right panels possess different signs, with the left panel suggesting that A_2 alleles increase the phenotype and the right panel the opposite. The regression lines are so different, because they minimize the sums of squared distances to the asterisks weighted by the relative frequencies of the three values of S. The weights are $(0.25, 0.5, 0.25)$ and $(0.01, 0.18, 0.81)$, the left and right panels, respectively.

The implication is that a linear regression of the traits of a random sample from a population on the number of A_2 alleles can be very misleading about the causal effects of the alleles. If A_1 is rare, then the regression will be driven by the people with alleles A_2.

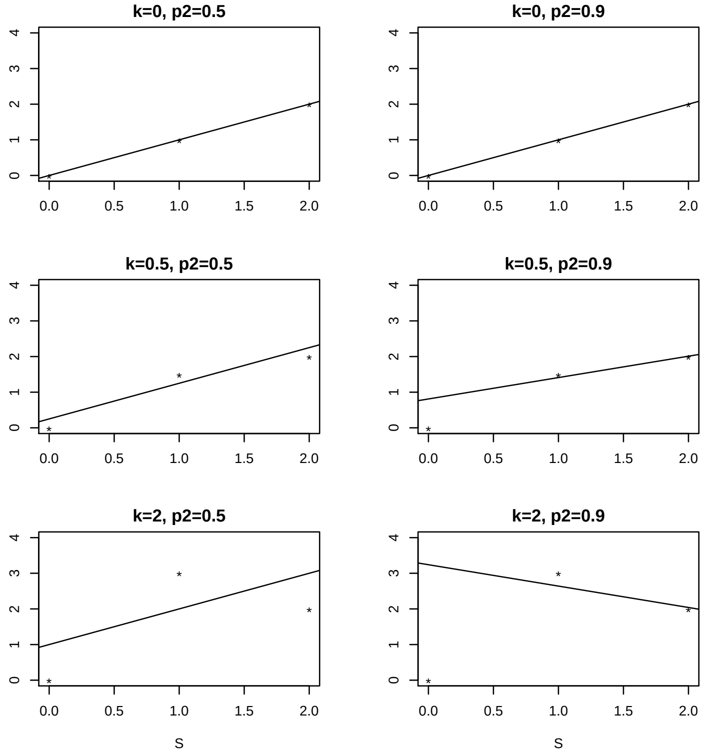

Figure 6.1. Regression of a monogenetic, biallelic trait (vertical axis) with alleles A_1, A_2 on the number of alleles A_2, for homozygous effect $a = 1$ and various values of dominance effect k and frequency of allele A_2. The vertical height of the three dots indicate the causal effect $a(1+(p_1-p_2)k)S$; for clarity the dots are plotted slightly higher than their actual values. The population is assumed to be in Hardy-Weinberg equilibrium. The slope of the regression line is $a(1+(p_1-p_2)k)$.

6.5 EXERCISE. Show that any function $(G_P, G_M) \mapsto \alpha + \beta(1_{G_P=A_2} + 1_{G_M=A_2})$ can be written in the form $h(G_P) + h(G_M)$ for some function $h\colon \{A_1, A_2\} \to \mathbb{R}$.

* 6.1.2 Bigenetic Traits

Consider a trait that is completely determined by two genes. Write the ordered genotypes of the individual as

$$\begin{array}{c|c} G_{P,1} & G_{M,1} \\ G_{P,2} & G_{M,2} \end{array}.$$

Suppose that the trait can be expressed as $X = f(G_{P,1}, G_{P,2}, G_{M,1}, G_{M,2})$ for a function f that is invariant under permuting its first and third arguments or its second and fourth arguments. The Hoeffding composition of X takes the form

$$\begin{aligned}X = \mathrm{E}X &+ f_1(G_{P,1}) + f_1(G_{M,1}) + f_2(G_{P,2}) + f_2(G_{M,2})\\ &+ f_{11}(G_{P,1}, G_{M,1}) + f_{22}(G_{P,2}, G_{M,2})\\ &+ f_{12}(G_{P,1}, G_{P,2}) + f_{12}(G_{P,1}, G_{M,2}) + f_{12}(G_{M,1}, G_{P,2}) + f_{12}(G_{M,1}, G_{M,2})\\ &+ f_{112}(G_{P,1}, G_{M,1}, G_{P,2}) + f_{112}(G_{P,1}, G_{M,1}, G_{M,2})\\ &\qquad\qquad + f_{122}(G_{P,1}, G_{P,2}, G_{M,2}) + f_{122}(G_{M,1}, G_{P,2}, G_{M,2})\\ &+ f_{1122}(G_{P,1}, G_{M,1}, G_{P,2}, G_{M,2}).\end{aligned}$$

The second line gives the dominance terms, and the third the epistasis. If $F_1, F_1^1, F_1^2, F_2, F_2^1, F_2^2$ are independent random variables with F_1, F_1^1, F_1^2 distributed as a random allele at the first locus in the population and F_2, F_2^1, F_2^2 as a random allele at the second locus in the population, then

$$\begin{aligned}\operatorname{var} X = &\, 2\mathrm{E}f_1^2(F_1) + 2\mathrm{E}f_2^2(F_2)\\ &+ \mathrm{E}f_{11}^2(F_1^1, F_1^2) + \mathrm{E}f_{22}^2(F_2^1, F_2^2)\\ &+ 4\mathrm{E}f_{12}^2(F_1, F_2)\\ &+ 2\mathrm{E}f_{112}^2(F_1^1, F_1^2, F_2) + 2\mathrm{E}f_{122}^2(F_1, F_2^1, F_2^2)\\ &+ \mathrm{E}f_{1122}^2(F_1^1, F_1^2, F_2^1, F_2^2).\end{aligned}$$

It is customary to abbreviate the variances on the right, as separated by lines, as $\sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2$.

6.2 Covariance

The traits X^1 and X^2 of two individuals who belong to some pedigree will generally be positively dependent, because of shared genetic make-up. We shall assume their covariance under the assumption that the founders of the pedigree are randomly chosen from a population in equilibrium. The IBD-configuration of the alleles of the two individuals is the key to understanding the dependence.

Assume that the two traits can be written

$$\begin{aligned}X^1 &= f(G_{P,1}^1, \ldots, G_{P,k}^1, G_{M,1}^1, \ldots, G_{M,k}^1) = f(G_P^1, G_M^1),\\ X^2 &= f(G_{P,1}^2, \ldots, G_{P,k}^2, G_{M,1}^2, \ldots, G_{M,k}^2) = f(G_P^2, G_M^2),\end{aligned}$$

with $G_P^1, G_M^1, G_P^2, G_M^2$ four haplotypes of k loci, and f a function that is symmetric in its ith and $(i+k)$th arguments. The haplotypes $G_P^1, G_M^1, G_P^2, G_M^2$ are random vectors, whose randomness will be viewed as arising from two sources: the sampling of the founders, and the process of segregation of the founder alleles to the two relatives. The structure of the pedigree is assumed given and nonrandom.

If there are f founders and we write the founder haplotypes as $F^1, F^2, \ldots, F^{2f}$, then the two sources of randomness can be pictured as in Figure 6.2. The four

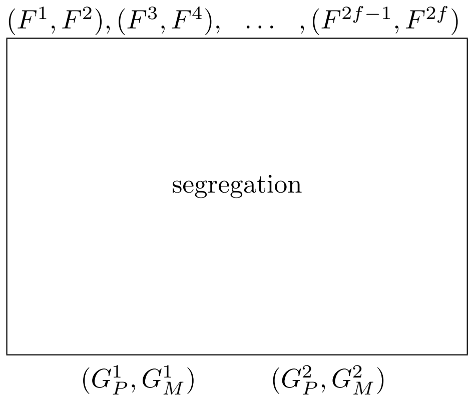
$(F^1, F^2), (F^3, F^4), \quad \dots \quad, (F^{2f-1}, F^{2f})$
segregation
(G_P^1, G_M^1)
(G_P^2, G_M^2)

Figure 6.2. Schematic representation of segregation.

haplotypes at the bottom of the box are recombinations and redistributions of the $2f$ haplotypes of the founders at the top of the box. We assume that the founders are chosen independently from a population that is in Hardy-Weinberg and linkage equilibrium, so that the $2f$ haplotypes $F^1, F^2, \dots, F^{2f}$ are i.i.d. random vectors of length k with independent marginals, $2kf$ independent alleles in total. As usual, we also assume that the segregation (inside the box) of the $2fk$ founder alleles is stochastically independent of the sampling of the founders. Segregation follows the branches of the pedigree (whose shape is assumed given) and the randomness consists of the choices of alleles passed on by the parents in meiosis, including recombination events.

Given what happens in the box the haplotypes $G_P^1, G_M^1, G_P^2, G_M^2$ of the two individuals of interest can be reconstituted from the founder haplotypes. We shall think of the four haplotypes as four vectors of length k with the k loci laid out along the horizontal axis, i.e. as four $(1 \times k)$-matrices, which can be joined into the $(4 \times k)$-matrix

$$\begin{array}{cccc} G_{P,1}^1 & G_{P,2}^1 & \cdots & G_{P,k}^1 \\ G_{M,1}^1 & G_{M,2}^1 & \cdots & G_{M,k}^1 \\ G_{P,1}^2 & G_{P,2}^2 & \cdots & G_{P,k}^2 \\ G_{M,1}^2 & G_{M,2}^2 & \cdots & G_{M,k}^2 \end{array}. \tag{6.6}$$

Numbering the $2f$ founders (arbitrarily) by the numbers $1, 2, \dots, 2f$, we can define for every locus i a segregation vector $V_i = (V_i^1, V_i^2, V_i^3, V_i^4)^T$ giving the founder labels (V_i^1, V_i^2) of the alleles of the first individual and the founder labels (V_i^3, V_i^4) of the alleles of the second individual at locus i (see Section 4.1). These vectors can be combined in a $(4 \times k)$-segregation matrix V whose entries correspond one-to-one to the entries of the matrix (6.6). This matrix completely describes the stochastic process inside the segregation box of Figure 6.2: given V the $(4 \times k)$-matrix of haplotypes $G = (G_P^1, G_M^1, G_P^2, G_M^2)^T$ in (6.6) are deterministic functions of the founder haplotypes $F^1, \dots, F^{2f}$. The distribution of the matrix G is best understood by conditioning on V:

$$P(G \in B) = \sum_v P(G \in B|\, V = v)P(V = v).$$

The number of different components $P(G \in B|\, V = v)$ in this finite mixture, and their weights $P(V = v)$ depend on the pedigree inside the box and the recombination properties between the loci.

Under the assumptions of Hardy-Weinberg and linkage equilibrium the components $P(G \in B|\, V = v)$ have a simple distribution. Given V the $(4 \times k)$-matrix (6.6) consists of particular founder alleles, where some founder alleles may occur multiple times. Because the founder alleles are independent and per locus identically distributed, the origins of the alleles are not important, but only the pattern of shared descending. Thus given V the joint law of the $(4 \times k)$-matrix G in (6.6) is the distribution of a $(4 \times k)$-matrix with:

(i) independent columns.
(ii) the four variables in the jth column are marginally distributed as an arbitrary allele for locus j in the population.
(iii) two variables in a given column are either identical (if the corresponding V's are equal) or independent.

It follows that the distribution of the $(4 \times k)$-matrix G can be completely described by the marginal distribution of the segregation vector V and the patterns of identical and independent variables in the separate columns as in (iii). For the latter we first study the case of a single-locus haplotype.

6.2.1 Single-locus Haplotypes

Consider the preceding in more detail for a single locus ($k = 1$), so that $F^1, \ldots, F^{2f}$ are i.i.d. univariate variables, and the segregation matrix V is the single vector $(V^1, V^2, V^3, V^4)^T$. The genotypes of the two individuals can be expressed in the founder alleles and segregation vectors as

$$(G_P^1, G_M^1) = (F^{V^1}, F^{V^2}) \quad \text{and} \quad (G_P^2, G_M^2) = (F^{V^3}, F^{V^4}).$$

The joint distribution of $(G_P^1, G_M^1, G_P^2, G_M^2)$ given $V = v$ is the distribution of the vector

$$(F^{v^1}, F^{v^2}, F^{v^3}, F^{v^4}).$$

There are $(2f)^4$ possible values v of V, and it suffices to determine the distribution of the vector in the display for every of them. Actually, as the founder alleles are i.i.d. random variables, the latter distribution does not depend on the exact values of $v^1, \ldots, v^4$, but only on the pattern of equal and unequal $v^1, \ldots, v^4$. If two coordinates are equal, we must insert the same founder allele, and otherwise independent copies. These patterns correspond to IBD-sharing, and there are (only) 15 possible IBD-configurations (or *identity states*) of the four alleles of the two individuals at a given locus. These are represented in Figure 6.3 by 15 graphs with four nodes. The two nodes at the top of a square represent the alleles G_P^1, G_M^1 of the first individual

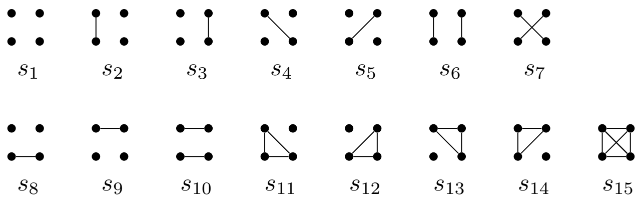

Figure 6.3. Identity states. Each of the 15 graphs represents an IBD-configuration of the alleles of two individuals at a given locus. The top two nodes of a graph represent the ordered genotype of the first individual, and the bottom two nodes the ordered genotype of the second individual. An edge indicates that the alleles are IBD.

and the two nodes at the bottom the alleles G_P^2, G_M^2 of the second individual; an edge indicates that the two alleles are IBD.

Configurations s_8 to s_{15} require that the pair of alleles of at least one of the two individuals is IBD (each graph has at least one horizontal edge). This can happen only if the pedigree is "inbred", a case that we shall usually exclude from consideration. Thus configurations s_1 to s_7, pictured in the upper row, are the more interesting ones. We shall refer to pedigrees for which identity states s_8 to s_{15} cannot occur as pedigrees without *inbreeding.* The possible distributions of the vector $(F^{v^1}, F^{v^2}, F^{v^3}, F^{v^4})$ for v belonging to one of the noninbred identity states are listed in Table 6.2.

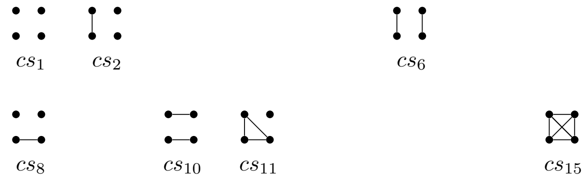

Figure 6.4. Condensed identity states. Each of the 7 graphs represents an IBD-configuration of the unordered sets of alleles of two individuals at a given locus. The top two nodes of a graph represent the unordered genotype of the first individual, and the bottom two nodes the unordered genotype of the second individual. An edge indicates that the alleles are IBD.

For many purposes the parental or maternal origin of the alleles is irrelevant, so that we can consider unordered gene pairs, and the two individuals under consideration can be swapped as well. The identity states can then be condensed to the collection shown in Figure 6.4. The condensed states cs_1, cs_2, and cs_6 are fully described by the IBD-sharing indicator N, which takes the value 0, 1 and 2 for the three states. Thus when excluding inbred states and ignoring parental and maternal origin we can describe the IBD-configuration completely by the variable N. The distributions of the possible vectors of unordered genotypes are given in Table 6.3, with the corresponding IBD-values given in the first column of the table.

v	$\mathcal{L}(G^1_P, G^1_M; G^2_P, G^2_M \,|\, V = v)$
s_1	$\mathcal{L}(F^1, F^2, F^3, F^4)$
s_2	$\mathcal{L}(F^1, F^2, F^1, F^3)$
s_3	$\mathcal{L}(F^1, F^2, F^3, F^2)$
s_4	$\mathcal{L}(F^1, F^2; F^3, F^1)$
s_5	$\mathcal{L}(F^1, F^2; F^2, F^4)$
s_6	$\mathcal{L}(F^1, F^2; F^1, F^2)$
s_7	$\mathcal{L}(F^1, F^2; F^2, F^1)$

Table 6.2. Conditional distribution of the ordered genotypes at a locus of two individuals given identity states s_1 to s_7. The variables F^1, F^2, F^3, F^4 are i.i.d. and distributed as a randomly chosen allele from the population.

N	v	$\mathcal{L}\big(\{G^1_P, G^1_M\}, \{G^2_P, G^2_M\} \,|\, V = v\big)$
0	cs_1	$\mathcal{L}\big(\{F^1, F^2\}, \{F^3, F^4\}\big)$
1	cs_2	$\mathcal{L}\big(\{F^1, F^2\}, \{F^1, F^3\}\big)$
2	cs_6	$\mathcal{L}\big(\{F^1, F^2\}, \{F^1, F^2\}\big)$

Table 6.3. Conditional distribution of the unordered genotypes at a locus of two individuals given condensed states cs_1, cs_2 and cs_7. The variables F^1, F^2, F^3, F^4 are i.i.d. and distributed as a randomly chosen allele from the population.

6.2.2 Multi-loci Haplotypes

Next consider haplotypes of k loci. Each locus is described by a set of founder alleles and an segregation vector, both specific to the locus. Because we assume linkage equilibrium, the founder alleles at different loci are always independent. It follows that the alleles of the two individuals at different loci are dependent only through the segregation vectors: the columns of the $(4 \times k)$ matrix $(G^1_P, G^1_M, G^2_P, G^2_M)^T$ are conditionally independent given the segregation matrix V. Because the marginal distributions of the columns depend only on the identity states, the joint distribution of the four haplotypes $G^1_P, G^1_M, G^2_P, G^2_M$ can be completely described by a (row) vector of identity states (one for each locus) and the marginal distributions of the alleles at the loci. Given the k-vector of identity states, the distribution of the $(4 \times k)$-matrix $(G^1_P, G^1_M, G^2_P, G^2_M)^T$ is equal to the independent combination of k columns distributed as the appropriate entry in Table 6.2 for the identity state for the corresponding locus. The k patterns of identity states are generally stochastically dependent.

If the k loci are unlinked, then both the founder alleles (outside the box in Figure 6.2) corresponding to different loci and the segregation process (inside the box) are independent across loci. Consequently, in this case the identity states for the loci are independent and the k columns of the $(4 \times k)$-matrix of haplotypes $G_P^1, G_M^1, G_P^2, G_M^2$ are independent, both conditionally and unconditionally. The unconditional distribution of this matrix can then be viewed as a mixture (with weights equal to the product of probabilities of the identity states), or as being constructed by joining k independent columns, each created by a single locus segregation process, as described in Section 6.2.1.

Because there are 15 possible identity states per locus, there are $(15)^k$ possible configurations (i.e. mixture components) of the joint distribution of $G_P^1, G_M^1, G_P^2, G_M^2$, too many to be listed. We can reduce the number of possibilities by considering unordered genotypes, but this still leaves many possibilities. In general it is not possible to describe the distribution by a vector of condensed identity states, one per locus, because condensing makes no difference between paternal and maternal origin and hence destroys haplotype information. However, for the purpose of deriving the covariance of traits that depend symmetrically on the two alleles at each locus, the haplotype information may be irrelevant and a reduction to condensed identity states may be possible. This is illustrated in the examples below.

There may be other simplifying structures as well. For instance, in a nuclear family (see Figure 5.3) the paternal alleles of the sibs are never IBD with the maternal alleles; for the cousins in Figure 7.1 there is a similar identification. In these cases certain values of v are impossible and the joint distribution of the unordered haplotypes $\{G_P^1, G_M^1\}, \{G_P^2, G_M^2\}$ can be completely described by the 3^k conditional joint distributions of the unordered haplotypes given the vector of IBD-values $(N_1, \ldots, N_k)$ at the k loci.

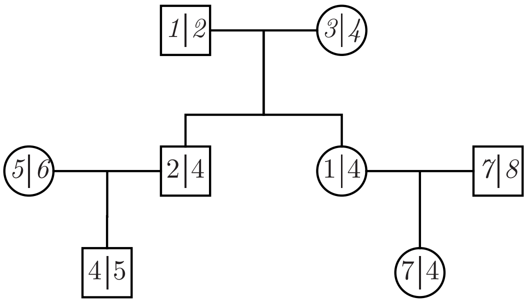

Figure 6.5. Cousins. Only one allele can be shared IBD.

6.2.3 General Rules

In view of the characterization of the distribution of the haplotypes as a mixture over the segregation matrix V, it is natural to compute the covariance $\mathrm{cov}(X^1, X^2)$ of the traits of the two individuals by conditioning on V. By the general rule on conditional covariances,

$$\operatorname{cov}(X^1,X^2)=\mathrm{E}\operatorname{cov}(X^1,X^2|\,V)+\operatorname{cov}\bigl(\mathrm{E}(X^1|\,V),\mathrm{E}(X^2|\,V)\bigr). \tag{6.7}$$

In the present case, in the absence of inbreeding, the covariance on the far right actually vanishes, as the conditional expectations $\mathrm{E}(X^i|\,V)$ do not depend on V. In fact, marginally the traits X^1 and X^2 are independent from the inheritance matrix, and distributed as the trait of an arbitrary person from the population.

6.8 Lemma. *Consider an individual in an arbitrary pedigree whose founders are drawn at random from a population that is in Hardy-Weinberg and linkage equilibrium. Then given $V=v$ for a v such that at no locus the individual's alleles at that locus are IBD, the individual's phenotype X is conditionally distributed as the phenotype of a random person from the population. In particular* $\mathrm{E}(X|\,V=v)=\mathrm{E}X$.

Proof. The segregation matrix completely determines how the founder alleles segregate to the individual. By Hardy-Weinberg and linkage equilibrium all founder alleles are independent. Given $V=v$ the individual's genotype consists of k pairs of copies of founder alleles. For v such that the alleles at no locus are IBD, these pairs consist of copies of two different founder alleles and hence are conditionally independent and distributed as a random allele for that locus. Given V the genotypes at different loci are independent, again because they are combinations of copies of founder alleles at different loci, which are independent. Thus the genotype of the individual is a set of independent pairs of alleles, each pair consisting of two independent alleles for that locus, the same as the genotpye of an arbitrary individual drawn from the population in Hardy-Weinberg and linkage equilibrium. The phenotype is a function of this genotype. ■

6.9 EXERCISE. Prove formula (6.7) for an arbitrary random vector (X^1,X^2,V).

In order to compute the conditional covariances $\operatorname{cov}(X^1,X^2|\,V)$ of the traits of the two individuals, we may decompose both variables X^1 and X^2 into their Hoeffding decompositions and calculate all cross-covariances between the terms of the decomposition conditioning on the segregation matrix V. This is not difficult, but can be tedious given the many different terms. A general rule is that in the absence of inbreeding cross-covariances between terms that do not depend on identical numbers of variables at each locus always vanish.

In particular, terms of different orders in the Hoeffding decompositions are conditionally uncorrelated.

6.10 Lemma. *Suppose that Y^1 and Y^2 are terms in the Hoeffding decompositions of $X^1=f(G^1_{P,1},G^1_{M,1},\ldots,G^1_{P,k},G^1_{M,k})$ and $X^2=f(G^2_{P,1},G^2_{M,1},\ldots,G^2_{P,k},G^2_{M,k})$ that depend on $(j^1_1,\ldots,j^1_k)$ and $(j^2_1,\ldots,j^2_k)$ variables at the loci $1,\ldots,k$ (where $j^i_l\in\{0,1,2\}$ for every l and i). If there is no inbreeding, then $(j^1_1,\ldots,j^1_k)\neq(j^2_1,\ldots,j^2_k)$ implies that* $\mathrm{E}(Y^1Y^2|\,V)=0$.

Proof. By the assumption of no inbreeding the alleles of a single individual at a given locus are not IBD. Therefore, given V the j_l^i variables of individual i at locus l are copies of different founder genes and hence independent random variables. If $j_l^1 < j_l^2$ for some locus l, then there must be a (copy of a) founder allele F in Y^2 that is not an argument of Y^1. We can write Y^i as a function g^i of $j_1^i+\cdots+j_k^i$ arguments. The joint conditional distribution of (Y^1, Y^2) given V is obtained by evaluating the functions (g^1, g^2) with as arguments the founder genes determined by V. Because Y^2 is a term of the Hoeffding decomposition of X^2, the expectation of the function g^2 with respect to a single argument relative to the marginal distribution of a random allele in the population vanishes, for any value of the other arguments. Let E_F denote the expectation with respect to F with the other alleles fixed. Because F does not appear in Y^1 we have $\mathrm{E}_F(Y^1Y^2|\, V) = Y^1\mathrm{E}_F(Y^2|\, V) = Y^1 0 = 0$. ■

6.2.4 Monogenetic Traits

Consider a trait that is completely determined by a single gene. Write the ordered genotypes of the two individuals as (G_P^1, G_M^1) and (G_P^2, G_M^2), respectively, and suppose that the traits can be expressed as $X^1 = f(G_P^1, G_M^1)$ and $X^2 = f(G_P^2, G_M^2)$ for a given function f that is symmetric in its arguments.

The Hoeffding decompositions of X^1 and X^2 take the forms:

$$\begin{aligned}
X^1 &= \mathrm{E}X^1 + f_1(G_P^1) + f_1(G_M^1) + f_{11}(G_P^1, G_M^1),\\
X^2 &= \mathrm{E}X^2 + f_1(G_P^2) + f_1(G_M^2) + f_{11}(G_P^2, G_M^2).
\end{aligned}$$

The functions f_1 and f_{11} are defined in Section 6.1.1. The linear parts $f_1(G_P^i) + f_1(G_M^i)$ and quadratic parts $f_{11}(G_P^i, G_M^i)$ are symmetric in their parental and maternal arguments G_P^i and G_M^i. Therefore, it suffices to consider condensed identity states. The three states that are possible in the absence of inbreeding are given in Table 6.2, together with the joint distribution of the genotypes, and correspond to 0, 1 or 2 alleles shared IBD by the two individuals. It follows that, for N the number of alleles shared IBD and F^1, F^2, F^3, F^4 i.i.d. variables distributed as a randomly chosen allele at the locus,

$$\begin{aligned}
\mathrm{cov}(X^1, X^2|\, N=0) &= \mathrm{E}\big(f_1(F^1) + f_1(F^2) + f_{11}(F^1, F^2)\big)\\
&\qquad\qquad \times \big(f_1(F^3) + f_1(F^4) + f_{11}(F^3, F^4)\big) = 0,\\
\mathrm{cov}(X^1, X^2|\, N=1) &= \mathrm{E}\big(f_1(F^1) + f_1(F^2) + f_{11}(F^1, F^2)\big)\\
&\qquad\qquad \times \big(f_1(F^1) + f_1(F^3) + f_{11}(F^1, F^3)\big)\\
&= \mathrm{E}f_1^2(F_1),\\
\mathrm{cov}(X^1, X^2|\, N=2) &= \mathrm{E}\big(f_1(F^1) + f_1(F^2) + f_{11}(F^1, F^2)\big)\\
&\qquad\qquad \times \big(f_1(F^1) + f_1(F^2) + f_{11}(F^1, F^2)\big)\\
&= 2\mathrm{E}f_1^2(F^1) + \mathrm{E}f_{11}^2(F^1, F^2).
\end{aligned}$$

In terms of the additive and dominance variances σ_A^2 and σ_D^2, as defined in Section 6.1.1, these three equations can be summarized as

$$\operatorname{cov}(X^1, X^2|\, N) = \tfrac{1}{2}\sigma_A^2 N + \sigma_D^2 1_{N=2}.$$

Consequently, with $\Theta = \frac{1}{4}\mathrm{E}N$ the kinship coefficient and $\Delta = P(N=2)$ the fraternity coefficient,

$$\operatorname{cov}(X^1, X^2) = 2\Theta\sigma_A^2 + \Delta\sigma_D^2.$$

6.11 Example (Sibs). For two sibs in a nuclear family $\Theta = \Delta = \frac{1}{4}$, and hence $\operatorname{cov}(X^1, X^2) = \frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2$. □

6.2.5 Additive and Dominance Covariance

Unless k is small the Hoeffding decompositions of the two traits X^1 and X^2 that depend on k loci have many terms. A common simplication is to assume that all terms except the linear term and the quadratic terms involving a single locus vanish, leaving only the additive and dominance variance terms. In other words, it is assumed that

$$\begin{aligned} X^1 &= \mathrm{E}X^1 + \sum_{j=1}^k \bigl(f_j(G_{P,j}^1) + f_j(G_{M,j}^1)\bigr) + \sum_{j=1}^k f_{jj}(G_{P,j}^1, G_{M,j}^1),\\ X^2 &= \mathrm{E}X^2 + \sum_{j=1}^k \bigl(f_j(G_{P,j}^2) + f_j(G_{M,j}^2)\bigr) + \sum_{j=1}^k f_{jj}(G_{P,j}^2, G_{M,j}^2). \end{aligned}$$

The variance of the traits can now be decomposed as

$$\operatorname{var} X^i = 2\sum_{j=1}^k \mathrm{E}f_j^2(F_j) + \sum_{j=1}^k \mathrm{E}f_{jj}^2(F_j^1, F_j^2) =: \sum_{j=1}^k \sigma_{A,j}^2 + \sum_{j=1}^k \sigma_{D,j}^2.$$

Here F_j, F_j^1, F_j^2 are i.i.d. and distributed as arbitrary alleles at locus j in the population.

Given the segregation matrix V the loci are independent and hence all conditional cross covariances between the terms in the decompositions of X^1 and X^2 depending on different loci vanish by Lemma 6.10 (or the explicit calculation in Section 6.2.4) vanish, and so all covariances between linear and quadratic terms. Only covariances between terms depending on a single loci remain, yielding

$$\begin{aligned} \operatorname{cov}(X^1, X^2|\, V) = \sum_{j=1}^k &\operatorname{cov}\bigl(f_j(G_{P,j}^1) + f_j(G_{M,j}^1), f_j(G_{P,j}^2) + f_j(G_{M,j}^2)|\, V\bigr)\\ &+ \sum_{j=1}^k \operatorname{cov}\bigl(f_{jj}(G_{P,j}^1, G_{M,j}^1), f_{jj}(G_{P,j}^2, G_{M,j}^2)|\, V\bigr). \end{aligned}$$

The conditional covariances on the right side are constant in V varying over a vector of condensed identity states. In the absence of inbreeding each such vector is completely described by the vector $(N_1,\ldots,N_k)$ of numbers of alleles shared IBD at loci $1,\ldots,k$. Exactly as in Section 6.2.4 we find that

$$\operatorname{cov}(X^1,X^2|\,V)=\tfrac{1}{2}\sum_{j=1}^{k}\sigma_{A,j}^2N_j+\sum_{j=1}^{k}\sigma_{D,j}^2 1_{N_j=2}.$$

The expected values of $\frac{1}{4}N_j$ and $1_{N_j=2}$ are the kinship and fraternity coefficients. Because these depend on the structure of the pedigree only, they are the same for all loci. Writing them as Θ and Δ, we find, with s_A^2 and σ_D^2 the total additive and dominance variance,

$$\operatorname{cov}(X^1,X^2)=2\Theta\sigma_A^2+\Delta\sigma_D^2.$$

6.2.6 Additive, Dominance and Epistatic Covariance

Consider the same situation as in Section 6.2.5, except this time assume that only the terms of order three and higher in the Hoeffding decomposition vanish, leaving the epistatic terms next to the linear and dominance terms. In other words, the traits are given by

$$\begin{aligned}X^1&=\mathrm{E}X^1+\sum_{j=1}^{k}\bigl(f_j(G_{P,j}^1)+f_j(G_{M,j}^1)\bigr)+\sum_{j=1}^{k}f_{jj}(G_{P,j}^1,G_{M,j}^1),\\&\quad+\sum\sum_{i<j}\bigl(f_{ij}(G_{P,i}^1,G_{P,j}^1)+f_{ij}(G_{M,i}^1,G_{M,j}^1)+f_{ij}(G_{P,i}^1,G_{M,j}^1)+f_{ij}(G_{M,i}^1,G_{P,j}^1)\bigr),\\X^2&=\mathrm{E}X^2+\sum_{j=1}^{k}\bigl(f_j(G_{P,j}^2)+f_j(G_{M,j}^2)\bigr)+\sum_{j=1}^{k}f_{jj}(G_{P,j}^2,G_{M,j}^2)\\&\quad+\sum\sum_{i<j}\bigl(f_{ij}(G_{P,i}^2,G_{P,j}^2)+f_{ij}(G_{M,i}^2,G_{M,j}^2)+f_{ij}(G_{P,i}^2,G_{M,j}^2)+f_{ij}(G_{M,i}^2,G_{P,j}^2)\bigr).\end{aligned}$$

The two expansions consist of the (constant) expectation plus three (random) sums. As always we assume absence of inbreeding.

In view of Lemma 6.10 the (conditional) covariances between non-corresponding terms vanish. The covariance contribution of the linear and dominance terms (the two sums in the first line) is exactly as in Section 6.2.5. In addition we obtain contributions of the form

$$\begin{aligned}\operatorname{cov}\bigl(&f_{ij}(G_{P,i}^1,G_{P,j}^1)+f_{ij}(G_{M,i}^1,G_{M,j}^1)+f_{ij}(G_{P,i}^1,G_{M,j}^1)+f_{ij}(G_{M,i}^1,G_{P,j}^1),\\&f_{ij}(G_{P,i}^2,G_{P,j}^2)+f_{ij}(G_{M,i}^2,G_{M,j}^2)+f_{ij}(G_{P,i}^2,G_{M,j}^2)+f_{ij}(G_{M,i}^2,G_{P,j}^2)|\,V\bigr).\end{aligned}$$

The two variables in this covariance are invariant under permutations per locus of the “P” and “M” symbols. The covariance is therefore invariant in matrices V that yield the same condensed identity states at loci i and j. It suffices to compute the covariance conditionally on the joint distribution of the IBD-indicators N_i and N_j at the two loci. Some thought shows that the sum over $i<j$ of the preceding display is equal to

$$\begin{aligned}\sum_{i<j}\sum\bigl(\mathrm{E}f_{ij}^2(F_i,F_j)\bigr)\bigl(1_{N_i=N_j=1}+2 1_{N_i=1,N_j=2}+2 1_{N_i=2,N_j=1}+4 1_{N_i=2,N_j=2}\bigr)\\ =\sum_{i<j}\sum\bigl(\mathrm{E}f_{ij}^2(F_i,F_j)\bigr)N_iN_j=:\tfrac14\sum_{i<j}\sum\sigma_{AA,ij}^2N_iN_j.\end{aligned}$$

The distribution of this term depends on the joint distribution of the IBD-indicators. For the expectation it suffices to know the joint distribution of two indicators, but the expectations $\mathrm{E}N_iN_j$ depend on the recombination fraction between the loci (i,j) as well as on the type of relationship between the individuals. We conclude that

$$\begin{aligned}\operatorname{cov}(X^1,X^2|\,V)&=\tfrac12\sum_{j=1}^k\sigma_{A,j}^2N_j+\sum_{j=1}^k\sigma_{D,j}^2 1_{N_j=2}+\tfrac14\sum_{i<j}\sum\sigma_{AA,ij}^2N_iN_j,\\ \operatorname{cov}(X^1,X^2)&=2\Theta\sigma_A^2+\Delta\sigma_D^2+\tfrac14\sum_{i<j}\sum\sigma_{AA,ij}^2\mathrm{E}N_iN_j.\end{aligned}$$

6.12 Example (Sibs). For sibs the joint distribution of a pair of IBD indicators is given in Table 4.3. In this case $\operatorname{cov}(N_i,N_j)=(1-\theta_{i,j})^2+\theta_{i,j}^2;=\psi_{i,j}$, and hence $\operatorname{cov}(X^1,X^2)=\frac12\sigma_A^2/2+\frac14\sigma_D^2+\sum\sum_{i<j}\sigma_{AA,ij}^2(\frac14\psi_{i,j}^2+1)$. □

6.2.7 Bigenetic Traits

Consider a trait that is completely determined by two genes. Write the ordered genotypes of the two individuals as

$$\begin{array}{c|c}G_{P,1}^1&G_{M,1}^1\\ G_{P,2}^1&G_{M,2}^1\end{array},\qquad\begin{array}{c|c}G_{P,1}^2&G_{M,1}^2\\ G_{P,2}^2&G_{M,2}^2\end{array}.$$

Suppose that the traits can be expressed as $X^i=f(G_{P,1}^i,G_{M,1}^i,G_{P,2}^i,G_{M,2}^i)$ for a function f that is invariant under permuting its first and second arguments or its third and fourth arguments. The Hoeffding decomposition of a trait of this type is given in Section 6.1.2, and displayed as consisting of a constant term plus five other terms. As before, we assume that the pedigree is not inbred. In view of Lemma 6.10 the (conditional) covariance between X^1 and X^2 can be obtained as the sum of covariances between the five terms. The covariances between the linear and quadratic terms are exactly as in Section 6.2.6.

The conditional covariance between the third order terms is given by

$$\begin{aligned}
\operatorname{cov}\bigl(&f_{112}(G^1_{P,1},G^1_{M,1},G^1_{P,2})+f_{112}(G^1_{P,1},G^1_{M,1},G^1_{M,2})\\
&\qquad+f_{122}(G^1_{P,1},G^1_{P,2},G^1_{M,2})+f_{122}(G^1_{M,1},G^1_{P,2},G^1_{M,2}),\\
&f_{112}(G^2_{P,1},G^2_{M,1},G^2_{P,2})+f_{112}(G^2_{P,1},G^2_{M,1},G^2_{M,2})\\
&\qquad+f_{122}(G^2_{P,1},G^2_{P,2},G^2_{M,2})+f_{122}(G^2_{M,1},G^2_{P,2},G^2_{M,2})|\,V\bigr)\\
&=1_{N_1=2}(1_{N_2=1}+21_{N_2=2})\mathrm{E}f^2_{112}(F^1_1,F^2_1,F_2)\\
&\qquad+1_{N_2=2}(1_{N_1=1}+21_{N_1=2})\mathrm{E}f^2_{122}(F_1,F^1_2,F^2_2)\\
&=1_{N_1=2}N_2\mathrm{E}f^2_{112}(F^1_1,F^2_1,F_2)+1_{N_2=2}N_1\mathrm{E}f^2_{122}(F_1,F^1_2,F^2_2).
\end{aligned}$$

The variances appearing on the right can be denoted by $\frac12\sigma^2_{DA,112}$ and $\frac12\sigma^2_{AD,122}$.

The conditional covariance contributed by the fourth order term is

$$\begin{aligned}
\operatorname{cov}\bigl(&f_{1122}(G^1_{P,1},G^1_{M,1},G^1_{P,2},G^1_{M,2}),f_{1122}(G^1_{P,1},G^1_{M,1},G^1_{P,2},G^1_{M,2})|\,V\bigr)\\
&\qquad=1_{N_1=N_2=2}\mathrm{E}f^2_{1122}(F^1_1,F^2_1,F^1_2,F^2_2).
\end{aligned}$$

The variance that appears on the right is denoted by σ^2_{DD}.

Taking all terms together we obtain the decomposition

$$\begin{aligned}
\operatorname{cov}(X^1,X^2|\,V)&=\tfrac12\sum_{j=1}^2\sigma^2_{A,j}N_j+\sum_{j=1}^2\sigma^2_{D,j}1_{N_j=2}+\tfrac14\sigma^2_{AA}N_1N_2\\
&\qquad+\tfrac12\sigma^2_{DA,112}1_{N_1=2}N_2+\tfrac12\sigma^2_{AD,122}1_{N_2=2}N_1+1_{N_1=N_2=2}\sigma^2_{DD}.
\end{aligned}$$

6.13 Example (Sibs). The joint distribution of the IBD-values N_1 and N_2 of two sibs in a nuclear family is given in Table 4.3. It follows that $\mathrm{E}N_j=1$, $P(N_j=2)=\frac14$, $\mathrm{E}N_1N_2=\psi+1$, $\mathrm{E}1_{N_1=2}N_2=\frac12\psi$, $\mathrm{E}1_{N_1=N_2=2}=\frac14\psi^2$, so that $\operatorname{cov}(X^1,X^2)=\frac12\sigma^2_A+\frac14\sigma^2_D+\frac14(\psi+1)\sigma^2_{AA}+\frac14\psi\sigma^2_{DA,112}+\frac14\psi\sigma^2_{AD,122}+\frac14\psi^2\sigma^2_{DD}$. □

7
Heritability

In this chapter we first extend the model for quantitative traits given in Chapter 6 to include environmental influences. This extension allows to discuss the extent by which a trait is genetically or environmentally determined. Such *biometrical analysis* has a long and controversial history (the *nature-nurture debate*), in which statistical techniques have played a big role. We discuss standard techniques to define and estimate heritability, and briefly discuss the philosophical significance of the results.

7.1 Environmental Influences

In Chapter 6 the trait X of a random individual from a population was written as a function $f(G)$ of the person's genes G. For most traits this is not realistic, because the trait will also depend on other factors, which can loosely be referred to as "environmental influences". The simplest model to include the environment is the *additive model*

$$X = f(G) + F,$$

for $f(G)$ the genetic factor as before, and F the environmental factor. The additive structure is special; independence of genes G and environment F makes the model even more special.

Additivity and independence are assumptions of a different nature. The independence of G and F refers to the randomness when sampling a person from the population. In our context sampling a person is equivalent to sampling a pair (G, F). Independence means that sampling a person is equivalent to sampling genes G and environment F independently and next combining these in a pair (G, F). This assumption would be violated, for instance, if the population consisted of two subpopulations with different genes, living in different environments. The independence does not by itself preclude causal interactions of genes and environment.

Causal interactions may come about after the genes G and environment F have been combined in an individual, for instance because certain genes may be disadvantageous in a certain environment. However, individuals must be alive to be sampled, and causal interactions may, through evolution, have created subpopulations (in the absence of random mating) that would lead to violation of independence in the sampling model.

The assumption of additivity has the direct causal implication that once a person's genes G and environment F are determined, the trait can be found by adding the influences $f(G)$ and F. In particular, this does exclude interaction.

The two assumptions also interact. The realism of the independence assumption depends on the population we sample from, but also on the way the trait is supposed to be formed once the pair (G, F) is determined. Perhaps, more realistically, genes and environment should be long vectors $(G_1, \ldots, G_k)$ and $(F_1, \ldots, F_l)$, not necessarily independent, and the trait a complicated function $f(G_1, \ldots, G_k, F_1, \ldots, F_l)$ of these vectors. However, we follow the tradition and work with the additive-independent model only.

When we consider two individuals, with traits X^1 and X^2, we must deal with two environmental influences. A common model is to assume that the environments of the two individuals can be decomposed in a *common environmental factor* C and *specific environmental factors* E^1 and E^2, and to assume that these factors are independent and act additively. This leads to the model

$$\begin{aligned} X^1 &= f(G^1) + C + E^1, \\ X^2 &= f(G^2) + C + E^2, \end{aligned} \tag{7.1}$$

where $(G^1, G^2), C, E^1, E^2$ are independent, and E^1, E^2 are identically distributed. The variables $f(G^i)$ are often abbreviated to A^i, giving $X^i = A^i + C + E^i$, which is known as the *A-C-E model*. Independence, as before, refers to the sampling of the individuals. Sampling a pair of individuals is equivalent to sampling their genes G^1 and G^2, a common environment C, and specific environments E^1 and E^2, all independently. Again these assumptions are an over-simplication of reality, but we adopt them anyway.

The specific environmental factors E^1 and E^2 also play the role of the ubiquitous "error variables" in statistical regression models: they make the models fit (better).

7.2 Heritability

For a trait X with decomposition $X = f(G) + F$, the fraction genetically determined variance or *heritability* is defined as

$$\frac{\operatorname{var} f(G)}{\operatorname{var} X} = \frac{\operatorname{var} f(G)}{\operatorname{var} f(G) + \operatorname{var} F}.$$

In this section we shall show that this number can be estimated from the observed trait values of relatives, under some assumptions. Often $\operatorname{var} f(G)$ is replaced in the numerator of the quotient by one of the approximations for $\operatorname{var} f(G)$ in Chapter 6. For instance, the additive approximation leads to the "fraction additively determined variance" or *additive heritability.*

A "heritability" of 60 % for a certain trait just means that the quotient of genetic over total variance is 0.6. It is difficult not to attach causal meaning to such a figure, but the definition has nothing causal about it. The variances in the quotient are variances measured in a given population, and are dependent on this population. For instance, the heritability of may traits is age-dependent, and can even be different for the same cohort of people at different ages. Also heritability of any trait in a genetically homogeneous is automically small, because most variation will be environmental. In the extreme case of a population consisting of a single genetic type, the heritability is zero.

These difficulties of interpretation come on top of the difficulties inherent in adopting the simplistic additive-independent model.

7.3 Biometrical Analysis

Consider estimating heritability from data. For simplicity we adopt a genetic model that incorporates additive and dominance effects only, as in Section 6.2.5. The analysis can easily be extended to genetic models involving more terms. Furthermore, we assume that environmental factors are added according to the additive-independent model described in Section 7.1.

Under these assumptions the variance and covariances of the traits X^1 and X^2 of two relatives can be decomposed as

$$\begin{aligned}\operatorname{var} X^1 = \operatorname{var} X^2 &= \sigma_A^2 + \sigma_D^2 + \sigma_C^2 + \sigma_E^2,\\ \operatorname{cov}(X^1, X^2) &= 2\Theta\sigma_A^2 + \Delta\sigma_D^2 + \sigma_C^2.\end{aligned}$$

The moments on the left sides of these equations can be estimated from observations on a random sample of traits of relatives. The kinship and fraternity coefficients Θ and Δ are numbers that can be computed from the type of relationship of the two relatives. Therefore, the resulting equations may be solved to find *moment estimators* for the unknown parameters $\sigma_A^2, \sigma_D^2, \sigma_C^2, \sigma_E^2$.

Because the moment equations are linear equations in four unknowns, we need at least four independent equations to identify the parameters. This may be achieved by sampling relatives of various kinds, leading to different pairs of kinship and fraternity coefficients Θ and Δ. Here we should be careful that the common and specific environment variances may vary with the different relatives also, and perhaps must be represented by multiple parameters. Another possibility to reduce the number of parameters is to assume that the dominance variance is zero. In the opposite direction, it is also possible to include additional variance terms, for instance the epistasis, as long as the kinship and fraternity coefficients of the individuals in our sample of observations are sufficiently varied.

An alternative for the method of moment is a likelihood based method. This might be applied to vectors $(X^1, \ldots, X^n)$ of more than two relatives at a time and employ their full joint distribution.

If the distribution of this vector is modelled jointly normal, then it suffices to specify variances and covariances, and the resulting estimates differ from the moment estimators mostly in the "pooling" of estimates from the various types of relatives. The means and variances of the variables $X^1, \ldots, X^n$ are a-priori assumed equal, and therefore the vector $X = (X^1, \ldots, X^n)$ is modelled to be $N_n(\mu 1, \sigma^2 H)$ distributed, where the $(n \times n)$-matrix H has entries

$$H_{i,j} = 2\Theta_{i,j} h_A^2 + \Delta_{i,j} h_D^2 + h_C^2.$$

The coefficients $\Theta_{i,j}$ and $\Delta_{i,j}$ are the kinship and fraternity coefficients of individuals i and j and can be computed from the structure of the pedigree. The unknown parameters in the model are the common mean value μ, the common variance σ^2, and the relative variances $h_A^2 = \sigma_A^2/\sigma^2$, $h_D^2 = \sigma_D^2/\sigma^2$ and $h_C^2 = \sigma_C^2/\sigma^2$. The parameter h_A^2 is the additive heritability and the sum $h_A^2 + h_D^2$ is the heritability. The maximum likelihood estimator maximizes the likelihood function, or equivalently minimizes the function (cf. Section 14.4)

$$(\mu, \sigma^2, h_A^2, h_D^2, h_C^2) \mapsto n \log \sigma^2 + \log \det H + \frac{1}{\sigma^2} \operatorname{tr}\Bigl(H^{-1}(X - \mu 1)(X - \mu 1)^T\Bigr).$$

When information on a sample of independent pedigrees is available, then this expression is of course summed over the pedigrees. It is shown in Lemma 14.37 that the maximum likelihood estimator for μ is simply the mean of all observations. The maximum likelihood estimators of the other parameters do not have explicit forms, but must be computed by a numerical algorithm.

7.2 Example (Twin design). It is not unreasonable to view monozygotic and dizygotic twins as comparable in all respects except genetic set-up. In particular, the environmental variances for monozygotic and dizygotic twin relatives could be modelled by the same parameter. Because monozygotic twins are genetically identical, their IBD-indicators are equal to 2 with probability 1, and hence their kinship and fraternity coefficients are 1/2 and 1. The genetic relationship of dizygotic twins does not differ from that of ordinary sibs, whence dizygotic twins have kinship and fraternity coefficients 1/4 and 1/4.

If we observe random samples of both monozygotic and dizygotic twins, then we may estimate the correlation between their traits by the sample correlation coefficients. It follows from the preceding that the population correlation coeffients ρ_{MZ} and ρ_{DZ} satisfy,

$$\rho_{MZ} = h_A^2 + h_D^2 + h_C^2.$$
$$\rho_{DZ} = \tfrac{1}{2} h_A^2 + \tfrac{1}{4} h_D^2 + h_C^2.$$

If we assume that the dominance variance σ_D^2 is 0, then this can be solved to give the fraction of genetic variance by the charming formula

$$h_A^2 = 2(\rho_{MZ} - \rho_{DZ}).$$

This can be estimated by replacing the correlation coefficients on the right by their estimates. (By definition the sample correlation coefficients are the sample covariance divided by the sample standard deviations. Because the standard deviations for the traits of the two individuals in a twin pair are assumed the same and also the same for monozygotic and dizygotic twins, it would be natural to adapt these estimates by employing a pooled variance estimator instead.)

The common environmental variance can be estimated from the correlations in a similar manner. The specific environmental variance can next be estimated using the decomposition of the variance σ^2. □

* 7.3.1 Categorical Traits

The variance decompositions of a trait obtained in Chapter 6 apply both to quantitative traits (which by definition can assume a continuum of possible values) and traits that can assume only finitely many values. However, for traits than can take only a few values, the decompositions are often viewed as less appropriate as a starting point for the analysis. It is common to assume that such a categorical trait is the result of a hidden, unobservable trait, often called a *liability*. The observed trait, for instance "diseased or not", "depression of a certain type or not depressed" would then correspond to the liability exceeding or not exceeding certain boundaries. The variance decomposition is applied to the liability.

Let (Y^1, Y^2) be the observed traits of a pair of relatives, assumed to assume values in a finite set $\{1, \ldots, m\}$. Let $\mathbb{R} = \cup_{j=1}^m I_j$ be a partition of the real line in intervals I_j and assume that there exist random variables X^1 and X^2 such that

$$Y^i = j \text{ if and only if } X^i \in I_j.$$

Thus the observed trait Y^i is a "discretization" of the liability X^i. If the intervals $I_1, \ldots, I_m$ are in the natural order, then high values of Y^i correspond to severe cases of the affection as measured in liability X^i.

The liabilities X^i are not observed, but "hidden"; they are also referred to as *latent variables*. Assume that these variables can be decomposed as

$$\begin{aligned} X^1 &= f(G^1) + C + E^1, \\ X^2 &= f(G^2) + C + E^2, \end{aligned}$$

with the usual independence assumptions between genetic and environmental factors. Then, with $A^i = f(G^i)$ the genetic component, by the independence of E^1 and E^2,

$$\begin{aligned} &P\bigl(Y^1 = y_1, Y^2 = y_2 |\, A^1, A^2, C\bigr) \\ &\quad = P\bigl(A^1 + C + E^1 \in I_{y_1}, A^2 + C + E^2 \in I_{y_2} |\, A^1, A^2, C\bigr) \\ &\quad = P^E(I_{y_1} - A^1 - C) P^E(I_{y_2} - A^2 - C). \end{aligned}$$

Here $P^E(I-x)=P(E+x\in I)$. It follows that the likelihood for observing the pair (Y^1,Y^2) can be written in the form

$$\int\!\!\int P^E(I_{Y^1}-a^1-c)P^E(I_{Y^2}-a^2-c)\,dP^{A^1,A^2}(a^1,a^2)\,dP^C(c).$$

The likelihood for observing a random sample of n of these pairs is the product of the likelihoods for the individual pairs. We can estimate the unknown parameters (mainly the variances of the variables A^1,A^2,C,E^1,E^2) using this likelihood.

Implementing the method of maximum likelihood or a Bayesian method is not trivial, but numerically feasible.

* 7.4 Regression to the Mean

The relationship between the traits of parents and their children was investigated at the end of the nineteenth century by Francis Galton. By numerical analysis of data from a large sample of parent-child pairs he concluded that the traits of children were closer to the mean of the sample than the parents' traits, whence the concept of *regression to the mean* was born, or in Galton's words "regression towards mediocrity". In his statistical analysis Galton was concerned with stature, but "mediocrity" receives a much more coloured significance if applied to traits as intelligence. Galton himself had earlier written on the heredity of "genius" and "talent", and was a founding father of *eugenics*, *"the science which deals with all influences that improve the inborn qualities of a race; also with those that develop them to the utmost advantage"*.[‡]

Galton's explanation of the phenomenon was a purely genetic one and was based on the idea that a child inherits part of his trait from its parents and the other part from its earlier ancestors. Regression to the mean would result from the fact the latter are more numerous and varied and more like random persons from the population the further they go back in the genealogy. This link of a child to its ancestors beyond its parents appears difficult to uphold. In fact, regression to the mean is purely statistical and the result of sampling from a population rather than causally determined. Imagine that trait values of individuals in a population are determined by systematic causes (including genes) and random causes. Then if a random parent is sampled from a population and happens to have a high trait value, it is likely that the random contribution to his trait is high as well. Because this is not inherited by the child, on the average the child will have a lower trait value.

We can obtain insight in this qualitative argument by the formulas for variance decompositions. If X^1 and X^2 are the traits of a parent and a child, then in the

[‡] Francis Galton, (1904). Eugenics: Its definition, scope, and aims, *The American Journal of Sociology 10(1).*

model (7.1) with the genetic factors given by dominance and epistatis terms

$$\operatorname{var} X^i = \sigma_A^2 + \sigma_D^2 + \sum\sum_{i<j} \sigma_{AA,ij}^2 + \sigma_C^2 + \sigma_E^2,$$

$$\operatorname{cov}(X^1, X^2) = \tfrac{1}{2}\sigma_A^2 + \tfrac{1}{4}\sum\sum_{i<j} \sigma_{AA,ij}^2 + \sigma_C^2.$$

In the second formula we use that a parent and child share exact one allele IBD, so that the kinship and fraternity coefficients are $\Theta = 1/4$ and $\Delta = 0$. The dominance variance does not appear in the covariance, because the two alleles at one locus come from different parents, which are assumed to be independently sampled from the population. The additive variance and epistasis do appear, but with a reduced coefficient. If $h_B^2 = \sigma_B^2/\sigma^2$ for each of the subscripts B and σ^2 the total variance $\sigma^2 = \operatorname{var} X^i$, then the correlation of the two traits is

$$\rho(X^1, X^2) = \tfrac{1}{2}h_A^2 + \tfrac{1}{4}\sum\sum_{i<j} h_{AA,ij}^2 + h_C^2.$$

This quantity is relevant to the explanation of the child's trait X^2 by the parent's trait X^1. In particular, the (population) least squares regression line is given by

$$\frac{x^2 - \mu}{\sigma} = \rho(X^1, X^2)\frac{x^1 - \mu}{\sigma}.$$

The regression to the mean is the phenomenon that the correlation coefficient $\rho(X^1, X^2)$ is smaller than 1. The calculations show that the latter is caused by the absence in the covariance of the specific environmental variance σ_E^2 (which is the purely random part), the absence of the dominance variance σ_D^2 (which is systematic, but unexplainable using one parent only), but also by the reduction of the contributions of additive variance and epistasis. Regression to the mean will occur even for a completely hereditary trait (i.e. $\sigma_C^2 = \sigma_E^2 = 0$). It will be stronger if the dominance variance is large.

The preceding formulas are based on the standard assumptions in Chapter 6, including equilibrium, random mating, and the additive-independent decomposition (7.1). In this set-up the means and variances of the population of parents and children do not differ, and the population as a whole does not regress and does not shrink to its mean. This is consistent with Galton's numerical observations, and made Galton remark that *"it is more frequently the case that an exceptional man is the somewat exceptional son of rather mediocre parents, than the average son of very exceptional parents".*[♭]

7.5 Prevalence

[♭] Francis Galton, (1886). Regression Towards Mediocrity in Hereditary Stature. *Journal of the Anthropological Institute 15*, 246-263.

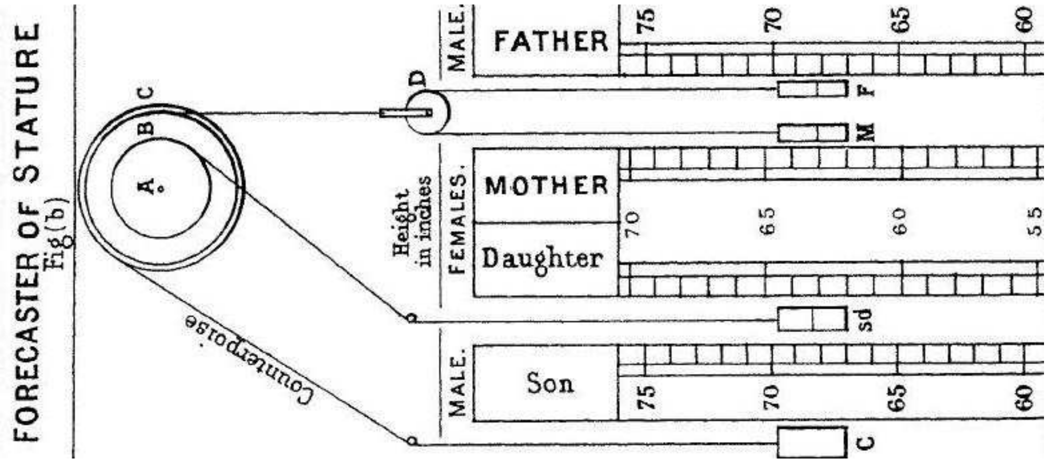

Figure 7.1. Galton's forecaster for the stature of child based on the statures of its parents, a graphical display of a regression function. From: Francis Galton, (1886). Regression Towards Mediocrity in Hereditary Stature. *Journal of the Anthropological Institute 15*, 246-263.

8
Quantitative Trait Loci

A quantitative trait is a phenotype that can assume a continuum of values, and a *quantitative trait locus* (or *QTL*) is a locus that causally influences the value of a quantitative trait. A typical quantitative trait is influenced by the genes at many loci as well as by the environment, which would explain its continuous nature.

In this chapter we study *nonparametric linkage methods* for finding quantitative trait loci. The underlying idea is the same as in Chapter 5, which was concerned with qualitative traits: the inheritance vectors of a pedigree at loci that are not linked to causal loci are independent of the trait. This principle can be operationalized in two ways. The first way is to model the conditional distribution of inheritance vectors (or IBD-values) given the trait, and test whether this really depends on the trait values. The second method is to model the conditional distribution of the trait given the inheritance vectors, and test whether this really depends on the latter. Whereas in Chapter 5 we followed the first route, in the present chapter we use the second. This is because it appears easier to model the distribution of a continuous trait given the discrete inheritance vectors than the other way around.

The models involved can be viewed as regression models for traits as dependent variables given IBD-sharing numbers as independent variables. Because the trait of a single individual is not dependent on IBD-sharing numbers, it is the dependence between traits of multiple individuals that must be regressed on the IBD-sharing numbers.

As dependence can be captured through covariance, the (conditional) covariance decompositions of Chapter 6 come in handy. These must of course be extended by adding environmental factors next to genetic vectors. We adopt the simplest possible model for these two influences, the additive model

$$X = f(G) + F,$$

where G, F are independent variables, corresponding to genetic make-up and environment, respectively. The genetic component of any individual is assumed independent of the environmental component of every other individual. Then the traits

X^1 and X^2 of two relatives with IBD-status N satisfy

$$\operatorname{cov}(X^1, X^2 |\, N) = \operatorname{cov}\bigl(f(G^1), f(G^2)|\, N\bigr) + \operatorname{cov}(F^1, F^2).$$

We use the models obtained in Chapter 6 for the first term on the right.

The environmental covariance $\operatorname{cov}(F^1, F^2)$ may be zero or nonzero, depending on the types of relatives involved. A simple customary model is that the two variables F^i can themselves be decomposed as $F^i = C + E^i$, where the variable C is common to the two relatives, the variables E^1, E^2 are identically distributed, and the three variables C, E^1, E^2 are independent. The variable C is said to reflect the *common environment*, whereas E^1, E^2 give the *specific environment.* In this model the environmental covariance $\operatorname{cov}(F^1, F^2) = \operatorname{var} C$ is precisely the variance of the common environmental factor. The variables C, E^1, E^2 are also assumed independent of the genetic factors.

The genetic factor $f(G)$ is often denoted by the letter A (presumably of "additive"), which leads to the decomposition $X = A + C + E$ of a trait. This is known as the *A-C-E model.*

8.1 Haseman-Elston Regression

A simple method, due to Haseman and Elston, performs linear regression of the squares $(X^1 - X^2)^2$ of the differences of the quantitative traits X^1 and X^2 of two relatives onto the IBD-counter N_u at a given locus u. The linear regression model for a single pair of relatives takes the form, wiht e an unobserved "error",

$$(X^1 - X^2)^2 = \alpha + \beta N_u + e. \tag{8.1}$$

If the coefficient β in this regression equation is significantly different from 0, then the locus u is implicated in the trait. More precisely, because we expect that a high value of IBD-sharing will cause the traits of the two relatives to be similar, a negative value of the regression coefficient β indicates that the locus is linked to the trait.

In practice we fit the regression model based on a random sample (X^{1i}, X^{2i}, N^i) of data on n pairs of relatives, and test the null hypothesis $H_0\colon \beta = 0$, for a large number of loci. We might use the usual t-test for this problem, based on the least squares estimator of β, or any other of our favourite testing procedures.

The regression equation can be understood quantitatively from the covariance decompositions in Chapter 6. If the pedigree to which the two individuals belong is not inbred and the population is in equilibrium, then the marginal distributions of X^1 and X^2 are equal and the two variables are marginally independent of IBD-status, by Lemma 6.8. In particular, the conditional means and variances of X^1 and X^2 given IBD-status are equal to the unconditional means and variances, which are equal for X^1 and X^2. It follows that

$$\mathrm{E}\bigl((X^1 - X^2)^2 |\, N_u\bigr) = \operatorname{var}(X^1 - X^2 |\, N_u) = 2 \operatorname{var} X^1 - 2 \operatorname{cov}\bigl(X^1, X^2 |\, N_u\bigr).$$

If a linear regression model for $\mathrm{cov}\bigl(X^1, X^2 |\, N_u\bigr)$ onto N_u is reasonable, then the Haseman-Elston procedure is justified. Because N_u assumes only three values, 0, 1 and 2, a quadratic regression model would certainly fit. A linear model should give a reasonable approximation. The slope of the regression of $\mathrm{cov}\bigl(X^1, X^2 |\, N_u\bigr)$ onto N_u should be positive, giving a negative slope for regression of $(X^1 - X^2)^2$ onto N_u. More precisely, if $\mathrm{cov}(X^1, X^2 |\, N_u) = \gamma + \delta N_u$, then the regression model (8.1) holds with $\beta = -2\delta$.

In practice the IBD-indicator N_u may not be observed. Then the regression is carried out on its conditional expectation $\mathrm{E}(N_u |\, M)$ given the observed (marker) data M instead.

Haseman-Elston regression is attractive for its simplicity, and for the fact that it requires hardly any assumptions. A drawback is that it reduces the data (X^1, X^2) to the differences $X^1 - X^2$, which may not capture all information about the dependence between X^1 and X^2 that is contained in their joint distribution. Furthermore, model assumptions (if they are reasonable) about the distribution of $X^1 - X^2$ may also help to increase the power of detecting QTLs.

8.2 Covariance Regression

Consider the joint conditional distribution of the trait values (X^1, X^2) of two relatives given the IBD-status N_U at a set U of loci of interest. If the pedigree is not inbred and the population is in equilibrium, then the marginal conditional distributions of X^1 and X^2 given N_U are equal and free of N_U, by Lemma 6.8. Thus a model of the joint conditional distribution of (X^1, X^2) given N_U should focus on the dependence structure.

If we assume that the conditional distribution of (X^1, X^2) given N_U is bivariate normal, then their complete distribution is fixed by the first and second order conditional moments. Because the mean vector and variances depend only on the marginal distributions, these quantities should be modelled independently of N_U. The conditional covariance $\mathrm{cov}(X^1, X^2 |\, N_U)$ is the only part of the model that captures the dependence. The results obtained in Chapter 6 suggest a wealth of models.

For instance, we may assume that the trait depends on k causal loci, and can be completely described by additive and dominance effects only, thus ignoring epistasis and interactions of three or more alleles. This leads to the formulas

$$\mathrm{var}(X^1 |\, N) = \mathrm{var}(X^2 |\, N) = \sum_{j=1}^k \sigma_{A,j}^2 + \sum_{j=1}^k \sigma_{D,j}^2 + \sigma_C^2 + \sigma_E^2, \tag{8.2}$$

$$\mathrm{cov}(X^1, X^2 |\, N) = \tfrac{1}{2}\sum_{j=1}^k \sigma_{A,j}^2 N_j + \sum_{j=1}^k \sigma_{D,j}^2 1_{N_j=2} + \sigma_C^2. \tag{8.3}$$

Here $N = (N_1, \ldots, N_k)$ is the vector of IBD-sharing indicators for the k causal loci, and σ_C^2 and σ_E^2 are the common and specific environment variances.

There are $2 + 2k$ unknown parameters in this display. These are identifiable from the distribution of (X^1, X^2, N), and can, in principle, be estimated from a sample of observations (X^{1i}, X^{2i}, N^i), for instance by using the moment equations

$$\begin{aligned} \sigma_C^2 &= \operatorname{cov}(X^1, X^2 \mid N = 0), \\ \tfrac{1}{2}\sigma_{A,j}^2 + \sigma_C^2 &= \operatorname{cov}(X^1, X^2 \mid N_j = 1; N_u = 0, u \neq j), \\ \sigma_{A,j}^2 + \sigma_{D,j}^2 + \sigma_C^2 &= \operatorname{cov}(X^1, X^2 \mid N_j = 2; N_u = 0, u \neq j). \end{aligned}$$

We can estimate the left sides by replacing the right sides by the appropriate empirical estimates, and next solve for the parameters on the left sides from top to bottom. (This method of estimation is only mentioned as a simple proof of identifiability.)

In practice, we do not know the number and locations of the causal loci, and typically we observe the IBD-status only at marker loci. If it is suspected that the number of causal loci is high, it may also be hard or impossible to fit a regression model that conditions on all such loci, as the resulting estimates will have high uncertainty margins. For instance, the method mentioned in the preceding paragraph cannot be implemented in practice unless we have a large number of observations: as the vector N can assume 3^k different values, the sets of observations with $(N_j = 1; N_u = 0, u \neq j)$ will be small or even empty, and the corresponding empirical estimators imprecise. Another difficulty is that it is not a-priori clear which loci to involve in the regression model. Typically one would like to scan the genome (or subregions) for loci, rather than test the effects of a few specific loci. If k is not small (maybe even $k = 2$ is already to be considered large), then there are too many sets of k loci to be taken into consideration.

For these reasons we simplify and model the conditional distribution of (X^1, X^2) given IBD-status N_u at a single marker locus u. Under the assumption of conditional normality, we only need to model the conditional covariance of (X^1, X^2) given N_u. This can be derived from the conditional covariance given the IBD-indicators $N = (N_1, \ldots, N_k)$ at the causal loci, as follows. If V is the segregation matrix at the causal loci, then the conditional mean $\mathrm{E}(X_i \mid N_u, V)$ is equal to the unconditional mean $\mathrm{E}X^i$ and hence nonrandom, by Lemma 6.8. Therefore, by the general conditioning rule for covariances (see Problem 6.9),

$$\operatorname{cov}(X^1, X^2 \mid N_u) = \mathrm{E}\bigl(\operatorname{cov}(X^1, X^2 \mid N_u, V) \mid N_u\bigr) = \mathrm{E}\bigl(\operatorname{cov}(X^1, X^2 \mid V) \mid N_u\bigr),$$

because (X^1, X^2) and N_u are conditionally independent given V, by Theorem 4.8. Using the model that incorporates additive and dominance terms given in (8.3), we find

$$\operatorname{cov}(X^1, X^2 \mid N_u) = \tfrac{1}{2}\sum_{j=1}^k \sigma_{A,j}^2 \mathrm{E}(N_j \mid N_u) + \sum_{j=1}^k \sigma_{D,j}^2 P(N_j = 2 \mid N_u) + \sigma_C^2.$$

If the locus j is not linked to the locus under investigation u, then N_j and N_u are independent and the conditional expectations $\mathrm{E}(N_j \mid N_u)$ and $P(N_j = 2 \mid N_u)$ reduce to four times the kinship coefficient $\Theta = \mathrm{E}\frac{1}{4}N_j$ and the fraternity coefficient

$\Delta = P(N_j = 2)$, corresponding to the relationship of the individuals. In particular, if only one of the causal trait loci j, say $j = 1$, is linked to the locus of current interest u, then the preceding display reduces to

$$\begin{aligned}
&\operatorname{cov}(X^1, X^2|\, N_u)\\
&\quad = \tfrac{1}{2}\sigma^2_{A,1}\mathrm{E}(N_1|\, N_u) + \sigma^2_{D,1}P(N_1 = 2|\, N_u) + 2\Theta\sum_{j=2}^k \sigma^2_{A,j} + \Delta\sum_{j=2}^k \sigma^2_{D,j} + \sigma^2_C\\
&\quad = \tfrac{1}{2}\sigma^2_{A,1}\big(\mathrm{E}(N_1|\, N_u) - 4\Theta\big) + \sigma^2_{D,1}\big(P(N_1 = 2|\, N_u) - \Delta\big) + 2\Theta\sigma^2_A + \Delta\sigma^2_D + \sigma^2_C.
\end{aligned}$$

Here σ^2_A and σ^2_D are the total additive and dominance variances, respectively, which are expressed as sums in the right side of (8.2). In the last expression the terms involving N_u are centered at mean zero, by the definitions of Θ and Δ. The six variances $\sigma^2_A, \sigma^2_D, \sigma_{A,1}, \sigma^2_{D,1}, \sigma^2_C, \sigma^2_E$ in the preceding display and (8.2) are unknown parameters. The kinship and fraternity coefficients Θ and Δ can be computed from the positions of the two relatives in the pedigree. Similarly, the conditional expectations $\mathrm{E}(N_1|\, N_u)$ and $P(N_1 = 2|\, N_u)$ depend on the structure of the pedigree and the recombination fraction between the loci 1 and u.

8.4 Example (Sibs). The conditional joint distribution of the IBD-sharing indicator at two loci of sibs in a nuclear family is given in Table 4.3. By some algebra it can be derived from this table that

$$\begin{aligned}
\mathrm{E}(N_u|\, N_v) &= 1 + (N_u - 1)e^{-4|u-v|},\\
P(N_u = 2|\, N_v) &= \tfrac{1}{4} + \tfrac{1}{2}(N_u - 1)e^{-4|u-v|} + \tfrac{1}{2}\big((N_u - 1)^2 - \tfrac{1}{2}\big)e^{-8|u-v|}.
\end{aligned}$$

These formulas should of course be evaluated only for $N_u \in \{0, 1, 2\}$, and can be writtern in many different forms. Note that the terms involving N_u at the right side are centered at mean zero. □

8.5 EXERCISE. Show that the second equation in Example 8.4 can also be written in the form $P(N_u = 2|\, N_v) = \frac{1}{4} + \frac{1}{2}(N_u - 1)e^{-4|u-v|}(1 - e^{-4|u-v|}) + (1_{N_u=2} - \frac{1}{4})e^{-8|u-v|}$.

As we are mainly interested in the dependence of the covariance on N_u, it is helpful to lump parameters together. Because N_u takes on only three values, any function of N_u can be described by three parameters, e.g. in the form $\alpha' + \beta'(N_u - 4\Theta) + \gamma'(1_{N_u=2} - \Delta)$. A convenient parameterization is the model

$$\begin{aligned}
\mathrm{E}(X^1|\, N_u) &= \mathrm{E}(X^2|\, N_u) = \mu,\\
\operatorname{var}(X^1|\, N_u) &= \operatorname{var}(X^2|\, N_u) = \sigma^2, \qquad (8.6)\\
\operatorname{cov}(X^1, X^2|\, N_u) &= \sigma^2\big(\rho + \beta(N_u - 4\Theta) + \gamma(1_{N_u=2} - \Delta)\big).
\end{aligned}$$

Here μ and σ^2 are the mean and variance of X^1 and X^2, both conditional and unconditional, and $\rho = \rho(X^1, X^2)$ is their unconditional correlation. This model has 6 unknown parameters. However, beware that the parameter ρ now incorporates the kinship and fraternity coefficients and the common environmental variances, so that it should be taken differently for different types of relatives.

The conventional method of analysis is to proceed by assuming that given N_u the vector (X^1, X^2) is bivariate normally distributed with parameters satisfying the preceding equations. Linkage of a locus u to the disease is investigated by testing the null hypothesis $H_0\colon \beta = \gamma = 0$ that the parameters in the conditional covariance involving N_u are zero. Rejection of the null hypothesis indicates linkage of the locus u to the disease. Because the alternative hypothesis can be taken that these parameters are negative, we may use a one-sided test.

To perform the test for multiple loci u the score test has the practical advantage that it suffices to compute the maximum likelihood estimator under the null hypothesis only once, as the null hypothesis is the same for every u. This estimator and the observed value of N_u are plugged in to a fixed function, involving the scores for β and γ. In contrast, the likelihood ratio statistic requires computation of the full maximum likelihood estimator for every locus u.

8.7 Example (Bivariate Normal Distribution). As a concrete example of the preceding inference, consider the model where a pair of traits (X^1, X^2) are assumed to be normal with mean $\mu = 0$, variance $\sigma^2 = 1$, and with the dominance variance a-priori assumed to be zero: $\gamma = 0$. Thus a typical observation is a triple (X^1, X^2, N) from the model described by:

(i) Given N the pair (X^1, X^2) is bivariate Gaussian with mean zero, variances 1 and covariance $\rho(N) = \rho + \beta(N - 4\Theta)$. The parameters ρ and β are unknown.

(ii) The variable N is 0, 1 or 2 with $\mathrm{E}N = 4\Theta$ and $\operatorname{var} N = \sigma_N^2$.

The log likelihood for this model is, up to a constant,

$$(\rho, \beta) \mapsto -\tfrac{1}{2}\log\bigl(1 - \rho^2(N)\bigr) - \tfrac{1}{2}\frac{(X^1)^2 + (X^2)^2 - 2\rho(N)X^1X^2}{1 - \rho^2(N)}.$$

The score vector for the parameter (ρ, β) is

$$\begin{pmatrix} 1 \\ N - 4\Theta \end{pmatrix} \Bigl[\frac{\rho(N)}{1 - \rho^2(N)} + \frac{X^1X^2}{1 - \rho^2(N)} - \frac{\bigl((X^1)^2 + (X^2)^2 - 2\rho(N)X^1X^2\bigr)\rho(N)}{(1 - \rho^2(N))^2}\Bigr].$$

We can rewrite the score in the form

$$\begin{pmatrix} 1 \\ N - 4\Theta \end{pmatrix} S_{\rho,\beta}(X^1, X^2, N),$$

for

$$S_{\rho,\beta}(X^1, X^2, N) = \frac{\bigl(X^1X^2 - \rho(N)\bigr)\bigl(1 + \rho^2(N)\bigr) - \rho(N)((X^1)^2 + (X^2)^2 - 2)}{(1 - \rho^2(N))^2}.$$

(This is elementary algebra. Alternatively, we can use that the score has conditional mean zero given N combined with the facts that $\mathrm{E}(X^1X^2 \,|\, N) = \rho(N)$ and

$\mathrm{E}((X^1)^2|\,N) = \mathrm{E}((X^2)^2|\,N) = 1$.) The Fisher information matrix in one observation is equal to

$$\mathrm{E}_{\rho,\beta}\begin{pmatrix}1\\ N-4\Theta\end{pmatrix}\begin{pmatrix}1\\ N-4\Theta\end{pmatrix}^T S^2_{\rho,\beta}(X_1^1, X_1^2, N)$$

Under $\beta = 0$ we have that $\rho(N) = \rho$, whence the variables (X^1, X^2) and N are independent and $S_{\rho,0}(X^1, X^2, N)$ is a function of (X^1, X^2) only. Then the information matrix becomes

$$\begin{pmatrix}1 & 0\\ 0 & \sigma_N^2\end{pmatrix}\tau^2$$

for

$$\tau^2 = \mathrm{E}_{\rho,0}\Bigl(\frac{(X^1X^2-\rho)(1+\rho^2)-\rho((X^1)^2+(X^2)^2-2)}{(1-\rho^2)^2}\Bigr)^2.$$

The (one-sided) score test for $H_0\colon \beta = 0$ rejects the null hypothesis for large values of

$$\frac{1}{\sigma_N\hat\tau\sqrt{n}}\sum_{i=1}^n (N^i - 4\Theta)S_{\hat\rho,0}(X^{1i}, X^{2i}, N^i).$$

Here $\hat\tau$ is obtained by replacing the expectation in the definition of τ by an average over the sample, and ρ by its maximum likelihood estimator under the null hypothesis, which is the solution to the equation $\sum_{i=1}^n S_{\rho,0}(X^{1i}, X^{2i}, N^i) = 0$, i.e. the solution to

$$\frac{1}{n}\sum_{i=1}^n X^{1i}X^{2i} = \rho + \frac{\rho}{1+\rho^2}\Bigl(\frac{1}{n}\sum_{i=1}^n\bigl((X^{1i})^2 + (X^{2i})^2\bigr) - 2\Bigr).$$

Because the term in brackets on the right is $O_P(1/\sqrt{n})$, the maximum likelihood estimator is asymptotically equivalent to the estimator $n^{-1}\sum_{i=1}^n X^{1i}X^{2i}$, which in turn is asymptotically equivalent to the sample correlation coefficient.

In practice, it is rarely known a-priori that the mean and variance of the observations are 0 and 1. However, this situation is often simulated by replacing each observation (X^1, X^2) by its “z-score” $(X^1 - \hat\mu, X^2 - \hat\mu)/\hat\sigma$, for $\hat\mu$ and $\hat\sigma$ the mean and sample standard deviation of all observed traits. The analysis next proceeds as if these standardized observations are multivariate normal with mean vector 0 and variance 1. Actually, if the original observation is multivariate normal, then its standardized version is not. However, some justification for this practice follows from the fact that the score functions for μ and σ^2 in the model (8.6) are orthogonal to the score functions for β and γ if evaluated at the null hypothesis (see below). This can be shown to imply that elimination of the parameters μ and σ^2 by standardization leads to the same inference for large numbers of observations. □

Warning. The assumptions that the conditional distribution of (X^1, X^2) given N is bivariate normal and the conditional distribution of (X^1, X^2) given N_u are bivariate normal, are typically mathematically incompatible. For instance, if $N_u =$

N_1 is the first coordinate of N, then the second conditional distribution can be derived from the first by the formula

$$\mathcal{L}\big((X^1, X^2) |\, N_1\big) = \sum_{n_2,\ldots,n_k} \mathcal{L}\big((X^1, X^2) |\, N_1, N_2 = n_2, \ldots, N_k = n_k\big) \\ \times P(N_2 = n_2, \ldots, N_k = n_k |\, N_1).$$

If (X^1, X^2) given $N = n$ is bivariate normal for every n, then this formula shows that the distribution of (X^1, X^2) given N_1 is a finite mixture of bivariate normal distributions. The terms of the mixture have the same mean vectors, but typically covariance matrices that depend on $(n_2, \ldots, n_k)$. Such a mixture is not bivariate normal itself. The same argument applies to a general locus u. In practice one does not worry too much about this inconsistency, because it is thought that the methods used, although motivated by the normality assumption, are not very dependent on this assumption. Furthermore, the conditional normality given N_u, used in the preceding, may be the more natural one if there are many causal loci, as mixtures over many components might make the distribution more continuous.

8.2.1 General Pedigrees

It is easy to incorporate sets of more than two relatives in the analysis. The trait vector $X = (X^1, \ldots, X^n)$ of n relatives is assumed to be conditionally distributed according to a multivariate normal distribution. The unknown parameters in this distribution are exactly the means, variances and covariances, and hence are specified as in the case of bivariate trait vectors. Specifically, given N_u the vector $X = (X^1, \ldots, X^n)$ is modelled to be $N_n(\mu 1, \sigma^2\Sigma)$-distributed, where the matrix Σ has (i, j)th element

$$\Sigma^{i,j} = \rho^{i,j} + \beta(N_u^{i,j} - 4\Theta^{i,j}) + \gamma(1_{N_u^{i,j}=2} - \Delta^{i,j}). \tag{8.8}$$

Here $N_u^{i,j}$ is the number of alleles carried IBD by the relatives i and j, and the unconditional correlation $\rho^{i,j}$ and the kinship and fraternity coefficients $\Theta^{i,j}$ and $\Delta^{i,j}$ may be specific to the relationship of the individuals i and j. (For $i = j$ we set $N_u^{i,i} = 2 = 4\Theta^{i,i} = \Delta^{i.i}$ and $\rho^{i,i} = 1$, so that $\Sigma^{i,j} = 1$.) The parameter γ is often taken a-priori equal to zero, expressing an assumption of absence of dominance. If $N_u^{i,j}$ is not observed, then it is replaced by its conditional expectation given the data.

For simplicity we shall take the correlations $\rho^{i,j}$ for $i \neq j$ to be equal to a single parameter ρ in the following. To test the null hypothesis $H_0\colon \beta = \gamma = 0$ we could use the score test. The score functions for the parameters μ and σ^2, are given by (see Section 14.4):

$$\frac{1}{\sigma^2} 1^T \Sigma^{-1}(X - \mu 1)$$

$$-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(X - \mu 1)^T \Sigma^{-1}(X - \mu 1).$$

The score functions for the parameters $\rho^{i,j} \equiv \rho$, β and γ all take the general form

$$(8.9) \qquad -\frac{1}{2\sigma^2}\operatorname{tr}\bigl(\Sigma^{-1}\dot\Sigma\bigr) + \frac{1}{2\sigma^2}(X-\mu 1)^T\Sigma^{-1}\dot\Sigma\Sigma^{-1}(X-\mu 1),$$

where for the three parameters the matrix $\dot\Sigma$ must be taken equal to the three matrices

$$\bigl(1_{i\neq j}\bigr), \qquad \bigl((N_u^{i,j}-4\Theta^{i,j})1_{i\neq j}\bigr), \qquad \bigl((1_{N_u^{i,j}=2}-\Delta^{i,j})1_{i\neq j}\bigr),$$

respectively. To test the null hypothesis $H_0{:}\,\beta=\gamma=0$ the parameter $(\mu,\sigma^2,\rho,\beta,\gamma)$ is replaced by its maximum likelihood estimator $(\hat\mu_0,\hat\sigma_0^2,\hat\rho_0,0,0)$ under the null hypothesis. In particular, the matrix Σ reduces to $\hat\Sigma_0=\bigl(1_{i=j}+\hat\rho_0 1_{i\neq j}\bigr)$, and is free of the IBD-indicators $N^{i,j}$.

The score test for $H_0{:}\,\beta=\gamma=0$ measures the deviation of the scores for β and γ from 0. Because the variables $N_u^{i,j}-4\Theta^{i,j}$ and $1_{N_u^{i,j}=2}-\Delta^{i,j}$ possess mean zero and are independent of X under the null hypothesis, the score functions for β and γ have conditional mean zero given X. (Note that $\mathrm{E}\dot\Sigma=0$ for the second and third form of $\dot\Sigma$ and (8.9) is linear in $\dot\Sigma$.) Combined with the fact that the other score functions are functions of X only, it follows that the score functions for β and γ are uncorrelated with score functions for μ,σ^2 and ρ. Thus the Fisher information matrix for the parameter $(\mu,\sigma^2,\rho,\beta,\gamma)$ has block structure for the partition in the parameters (μ,σ^2,ρ) and (β,γ), and hence its inverse is the block matrix with the inverses of the two blocks. For the score statistic, this means that the weighting matrix is simply the Fisher information matrix for the parameter (β,γ). See Example 14.29.

8.2.2 Extensions

The preceding calculations can be extended to the situation that the locus u is linked to more than one of the causal loci $1,\ldots,k$. It is also possible to include external covariate variables (e.g. sex or age) in the regression equation. In particular, the mean μ may be modelled as a linear regression $\beta^T Z$ on an observable covariate vector Z.

8.2.3 Unobserved IBD-Status

In practice the IBD-status at the locus u may not be fully observed, in particular when the genes at the locus have only few alleles and/or the pedigree is small. This problem is usually overcome by replacing the IBD-variable N_u in the regression equation by its conditional expectation $\mathrm{E}(N_u|\,M)$ given observed marker data, computed under the assumption of no linkage. This is analogous to the situation in Chapter 5.

8.2.4 Multiple Testing

When testing a large set of loci u for linkage the problem of multiple testing arises. This too is analogous to the situation in Chapter 5.

8.2.5 Power

As to be expected, the bigger the variances $\sigma^2_{A,1}$ or $\sigma^2_{D,1}$ contributed by the locus $u = 1$, the bigger the parameters β and γ, and the more power to detect that the null hypothesis is false.

8.2.6 Epistasis

So far we have assumed that the epistatic component in the covariance is zero. Including epistasis may make the model more realistic and permit to find interactions between the loci. In theory it is even possible that two loci might not have "main effects", but do have a joint effect.

To include epistasis we replace the right side of (8.3) by

$$\operatorname{cov}(X^1, X^2 |\, N) = \tfrac{1}{2}\sum_{j=1}^k \sigma^2_{A,j} N_j + \sum_{j=1}^k \sigma^2_{D,j} 1_{N_j=2} + \sigma^2_C + \tfrac{1}{4}\sum\sum_{i<j} \sigma^2_{AA,ij} N_i N_j.$$

We may now follow the arguments leading eventually to the model (8.6) or (8.8). However, in deriving model (8.6) we already saw that any function of a single IBD-indicator N_u can be described by three parameters, and all of the three were accounted for by additive and dominance variances. Thus adding epistasis will not lead to a different model for the conditional distribution of the trait vector given N_u.

As is intuitively clear, to find epistasis we must study the conditional law of the traits given pairs IBD-indicators.??

* 8.3 Copulas

Because quantitative traits are marginally independent of the IBD-values, the analysis in this chapter focuses on the dependence structure of traits within their joint distribution given the IBD-values. The formal way of separating marginal and joint probability distributions is through "copulas".

Consider an arbitrary random vector $(X^1, \ldots, X^n)$, and denote its marginal cumulative distribution functions by $F^1, \ldots, F^n$ (i.e. $F^i(x) = P(X^i \leq x)$). The *copula* corresponding to the joint distribution of $(X^1, \ldots, X^n)$ is defined as the joint distribution of the random vector $\bigl(F^1(X^1), \ldots, F^n(X^n)\bigr)$. The corresponding multivariate cumulative distribution function is the function C defined by

$$C(u^1, \ldots, u^n) = P\bigl(F^1(X^1) \leq u^1, \ldots, F^n(X^n) \leq u^n\bigr).$$

Because each function F^i takes values in $[0, 1]$, the copula is a probability distribution on the unit cube $[0, 1]^n$. By the "probability integral transformation" each of the random variables $F^i(X^i)$ possesses a uniform distribution on $[0, 1]$, provided that F^i is a continuous function. Thus provided that the distribution of the vector

$(X^1,\ldots,X^n)$ possesses no atoms, the corresponding copula is a probability distribution on the unit cube with uniform marginals. If the cumulative distribution functions are strictly increasing, then we can invert the preceding display to see that, for every $(x^1,\ldots,x^n)\in\mathbb{R}^n$,

$$P\bigl(X^1\le x^1,\ldots,X^n\le x^n\bigr)=C\bigl(F^1(x^1),\ldots,F^n(x^n)\bigr).$$

With some difficulty it can be proved that for *any* multivariate distribution there exists a distribution on $[0,1]^n$ with uniform marginals whose cumulative distribution function C satisfies the display. (If the marginal distribution functions are continuous, then C is unique.) This fact is known as *Sklar's theorem.*

It follows that the joint distribution of the vector $(X^1,\ldots,X^n)$ can be completely described by the marginal distribution functions $F^1,\ldots,F^n$ and the copula C. Here the copula contains all information about the dependence between the coordinates X^i, this information being absent from the marginal distributions.

8.10 EXERCISE. Show that we can write any distribution function F in the form $F=C\circ(\Phi,\ldots,\Phi)$, for Φ the standard normal distribution function and C some distribution function on $\mathbb{R}^n$. [Thus the transformation to uniform marginals in the definition of a copula is arbitrary. Any nice distribution would do.]

8.11 Example (Gaussian copula). The bivariate normal distribution is specified by a mean vector $\mu\in\mathbb{R}^2$ and a covariance matrix Σ, which is a positive-definite (2×2)-matrix. The means and variances are parameters of the marginal distributions and hence the corresponding copula must be free of them. It follows that the copula can be described by a single parameter (corresponding one-to-one to the off-diagonal element of Σ), which we can take to be the correlation.

The normal copula does not permit expression in elementary functions, but takes the form

$$C_\rho(u^1,u^2)=\int_{-\infty}^{\Phi^{-1}(u^1)}\int_{-\infty}^{\Phi^{-1}(u^2)}\frac{1}{2\pi\sqrt{1-\rho^2}}e^{-\frac12(x^2-2\rho xy+y^2)/(1-\rho^2)}\,dx\,dy.$$

This is just the probability that $P\bigl(\Phi(X^1)\le u^1,\Phi(X^2)\le u^2\bigr)$ for (X^1,X^2) bivariate normal with mean zero, variance one and correlation ρ. □

In our application we use copulas for the conditional distribution of (X^1,X^2) given N. The usual assumption that this conditional distribution is bivariate normal is equivalent to the assumptions that the marginal distributions of X^1 and X^2 (given N) are normal and that the copula corresponding to the joint conditional distribution (given N) is the normal copula. Both assumptions could be replaced by other assumptions.

For instance, if the traits are the time of onset of a disease, then normal distributions are not natural. We could substitute standard distributions from survival analysis for the marginals, and base the copula on a Cox model.

Rather than distributions of a specific form we could also use semiparametric or nonparametric models.

* 8.4 Frailty Models

Frailty models have been introduced in survival analysis to model the joint distribution of survival times. They can be applied in genetics by modelling the frailty as a sum of genetic and environmental factors.

Let (T^1, T^2) be two event times for a related pair of individuals (twins, sibs, parent-child, etc.). Let (Z^1, Z^2) be a corresponding pair of latent variables ("frailties") such that T^1 and T^2 are conditionally independent given (Z^1, Z^2) with cumulative hazard functions $t \mapsto Z^1\Lambda(t)$ and $t \mapsto Z^2\Lambda(t)$, respectively, for a given "baseline hazard function" Λ. In other words, under the assumption that the conditional distribution functions are continuous, the joint conditional survival function is given by

$$P(T^1 > t^1, T^2 > t^2 |\, Z^1, Z^2) = e^{-Z^1\Lambda(t^1)} e^{-Z^2\Lambda(t^2)}.$$

The unconditional survival function of (T^1, T^2) is the expectation of this expression with respect to (Z^1, Z^2).

Thus under the model the conditional hazard functions of T^1 and T^2 are proportional, with the quotient of the frailties as the proportionality constant.

The marginal (i.e. unconditional) survival functions of T^1 and T^2 are given by $t \mapsto \mathrm{E}e^{-Z^1\Lambda(t)}$ and $t \mapsto \mathrm{E}e^{-Z^2\Lambda(t)}$, respectively. These are identical if Z^1 and Z^2 possess the same marginal distribution. We complete the model by choosing a marginal distribution and a copula for the joint distribution of (Z^1, Z^2).

An attractive possibility is to choose the marginal distribution *infinitely divisible*. Infinitely divisible distributions correspond one-to-one with *Lévy processes*: continuous time processes $Y = (Y_t: t \geq 0)$ with stationary, independent increments and $Y_0 = 0$. The corresponding infinitely divisible distribution is the distribution of the variable Y_1. From the decomposition $Y_1 = Y_\rho + (Y_1 - Y_\rho)$, it follows that Y_1 is for every $0 \leq \rho \leq 1$ distributed as the sum of two independent random variables distributed as Y_ρ and $Y_{1-\rho}$, respectively. Given independent copies Y and $\tilde{Y}$ we now define frailty variables

$$\begin{aligned} Z^1 &= Y_\rho + (Y_1 - Y_\rho) = Y_1, \\ Z^2 &= Y_\rho + (\tilde{Y}_1 - \tilde{Y}_\rho). \end{aligned}$$

Then Z^1 and Z^2 possess the same marginal distribution, and have correlation

$$\rho(Z^1, Z^2) = \frac{\operatorname{var} Y_\rho}{\operatorname{var} Y_1} = \rho.$$

In order to obtain nonnegative frailties the distribution of Y_1 must be concentrated on $[0, \infty)$. The corresponding Lévy process then has nonincreasing sample paths, and is called a *subordinator*.

With the frailties as given in the preceding display the unconditional joint survival function is given by

$$\begin{aligned} P(T^1 > t^1, T^2 > t^2) &= \mathrm{E} e^{-Z^1\Lambda(t^1)} e^{-Z^2\Lambda(t^2)} \\ &= \mathrm{E} e^{-Y_\rho(\Lambda(t^1)+\Lambda(t^2))} \mathrm{E} e^{-(Y_1-Y_\rho)\Lambda(t^1)} \mathrm{E} e^{-(\tilde{Y}_1-\tilde{Y}_\rho)\Lambda(t^2)} \\ &= \psi\bigl(\Lambda(t^1)+\Lambda(t^2)\bigr)^\rho \psi\bigl(\Lambda(t^1)\bigr)^{1-\rho} \psi\bigl(\Lambda(t^2)\bigr)^{1-\rho}, \end{aligned}$$

where $\psi(u) = \mathrm{E} e^{-uY_1}$ is the Laplace transform of Y_1. In the last step we use the identity $\mathrm{E} e^{-uY_t} = \bigl(\mathrm{E} e^{-Y_1}\bigr)^t$, which follows from the independence and stationarity of the increments. Setting t^2 in this display equal to zero shows that the marginal conditional survival functions are given by

$$S(t) \colon= P(T^1 > t) = \psi\bigl(\Lambda(t)\bigr).$$

We can write the joint survival function in terms of this function by substituting $\Lambda = \psi^{-1} \circ S$ in the second last display.

8.12 Example (Gamma frailties). The Gamma distribution with parameters λ and 1 (so that the mean is λ and the variance $1/\lambda^2$) is infinitely divisible. Its Laplace transform is $\psi(u) = (1+u)^\lambda$. The corresponding joint survival function is given by

$$P(T^1 > t^1, T^2 > t^2) = \Bigl(\frac{1}{1+\Lambda(t^1)+\Lambda(t^2)}\Bigr)^{\lambda\rho} \Bigl(\frac{1}{1+\Lambda(t^1)}\Bigr)^{\lambda(1-\rho)} \Bigl(\frac{1}{1+\Lambda(t^2)}\Bigr)^{\lambda(1-\rho)}.$$

The marginal survival functions are

$$S(t) = \Bigl(\frac{1}{1+\Lambda(t)}\Bigr)^\lambda.$$

Solving $\Lambda(t)$ from this and substitution in the preceding display shows that

$$P(T^1 > t^1, T^2 > t^2) = \Bigl(\frac{1}{1+S(t^1)^{-1/\lambda}+S(t^2)^{-1/\lambda}}\Bigr)^{\lambda\rho} S(t^1)^{1-\rho} S(t^2)^{1-\rho}.$$

This reveals the copula connecting the marginals of T^1 and T^2 in their joint distribution. The parameter ρ has the interpretation of correlation between the underlying frailties, whereas $1/\lambda^2$ is the variance of a frailty. □

A "correlated frailty model" can now be turned into a model for the conditional distribution of (T^1, T^2) given IBD-status by assuming that (T^1, T^2) is conditionally independent of N_u given $(Z_1^{;}Z^2)$, and (Z^1, Z^2) given N_u possesses (conditional) correlation

$$\rho(Z_1, Z_2 |\, N_u) = \rho + \beta(N_u - 4\Theta) + \gamma(1_{N_u=2} - \Delta).$$

9
Association Analysis

Perhaps the simplest method to find genes that cause an affection would be to compare the full genomes of samples of affected and unaffected individuals. The loci of causal genes must be among the loci where the two samples are significantly different. The main drawback of this approach is the size of the genome. Sequencing the genome of large numbers of individuals is still unfeasible, and analyzing the resulting large numbers of data would encounter both computational and theoretical challenges.

Association analysis consists of comparing the genomes of cases and controls at selected markers rather than at every locus. The name "association" is explained by the fact that this partial strategy can be successful only if the measured loci are "correlated" or "associated" to the causal loci for the affection. A population was defined to be in "linkage equilibrium" if the alleles at different loci on a randomly chosen haplotype are independent. Because in this situation a marker would never be informative about any other locus than itself, the opposite, which is called *linkage disequilibrium*, is needed for association analysis. More precisely, we need linkage disequilibrium between the marker and the causal loci.

In Section 2.3 it was seen that, under random mating, any sequence of populations eventually reaches linkage equilibrium. Furthermore, possible disequilibrium between two loci separated by recombination fraction θ is reduced by a factor $(1-\theta)^k$ in k generations. For this reason the assumption of equilibrium seemed reasonable in many situations, in particular for loci that are far apart. However, it is not unreasonable to expect a causal locus for a disease to be in linkage disequilibrium with marker loci that are tightly linked to it. Imagine that a disease-causing gene was inserted in the genome of an individual (or a set of individuals) a number of generations in the past by one or more mutations, and that the current disease-carrying subpopulation are the offspring of this individual (or set of individuals). The mutation would have broken linkage equilibrium (if this existed), as the diseased individuals would carry not only the mutated gene, but also an exact copy of a small segment around the gene of the DNA of the individual who was first affected. After repeated rounds of random mating, these segments would have become smaller, and the population would eventually return to linkage equilibrium, because recombination events occur between the mutated and other loci in a random fashion. The form of the reduction factor $(1-\theta)^k$ shows that the return to equilibrium is very rapid for loci that are far apart on the genetic map. However, if the mutation originated not too many generations in the past, then it is likely that loci close to the disease mutation are (still) in linkage disequilibrium with the disease locus.

This reasoning suggests that cases and controls may indeed differ at marker loci that are close to causal loci, so that an association study may work. It also suggests that marker loci at some distance of causal loci will not be associated to the causal loci. Association studies turn this lack of information in distant markers into an advantage, by becoming a method for *fine-mapping* of genes: markers at some distance of a causal locus will automatically be ignored. A linkage study of the type discussed in Chapters 3 or 5 would pin down some region of the genome that is likely to contain a causal gene. Next an association study would reveal the location within the region at higher precision.

If tightly linked loci are indeed highly correlated in the population, then measuring and comparing the full genome will not only be unpractical, but will also not add much information above testing only selected loci. In particular, for *genome wide association studies*, which search the complete genome for causal genes, it should be enough to use marker loci in a grid such that *every* locus has high correlation (e.g. higher than 0.8) with some marker locus. The *HapMap project* (see `http://www.hapmap.org/`) is a large international effort to find such a set of markers, by studying the variety of haplotypes in the world population, and estimate linkage disequilibrium between them. It is thought that a set of 600 000 SNPs could be sufficient to represent the full genome. Experimental chip technology of 2008 permits measurements of up to 100 000 SNPs in one experiment, and association studies may be carried out on as many as 20 000 cases and controls. Thus many researchers believe that genome-wide association studies are the most promising method to find new genes in the near future. However, other researchers are less optimistic, and claim that large-scale association studies are a waste of effort and money.

9.1 Association and Linkage Disequilibrium

Two alleles A_i and B_j on two different loci are defined to be *associated* within a given population if the frequency h_{ij} of the haplotypes A_iB_j and the frequencies p_i and q_j of the alleles A_i and B_j satisfy

$$h_{ij} \neq p_i q_j.$$

In other words, if we randomly choose one of the two haplotypes from a random individual from the population, then the alleles at the two loci are not independent.

Apart from the insistence on two particular alleles, this definition of association is the opposite of linkage equilibrium ("dependent" rather than "independent"). Thus association is the same as linkage *dis*equilibrium. However, several authors object to this identification, and would define linkage disequilibrium as "allelic association that has not been broken up by recombination". In their view not every association is linkage disequilibrium, and they prefer the term *gametic phase disequilibrium* over "association".

This (confusing) refinement is motivated by the different ways in which association may arise, and their consequences for statistical and genetic analysis. If association arises through a mutation in some ancestor at some locus, which is next inherited by offspring together with a chromosomal segment containing the other locus, then this constitutes the linkage disequilibrium of the type we are interested in. However, association as defined in the first paragraph of this section is simply statistical correlation and may arise in many other ways. Natural selection may go against certain combinations of alleles, or statistical fluctuations from generation to generation may cause deviations, particularly in small populations. Perhaps the most important cause for association is population substructure (also called *admixture* or *stratification*) that goes against random mating. For instance, suppose that the population consists of two subpopulations, and that each subpopulation is in linkage equilibrium (perhaps due to many rounds of random mating within the subpopulations). Then alleles in the full population will be associated as soon as the marginal frequencies of alleles in the subpopulations are different. This is nothing but an instance of the well-known *Simpson's paradox*, according to which two variables may be conditionally independent given a third variable, but not unconditionally independent.

To quantify this notion, consider two populations such that allele A has frequencies p^1 and p^2, allele B has frequencies q^1 and q^2, and haplotype AB has frequencies h^1 and h^2 in the populations. Let the two populations have relative sizes λ and $1-\lambda$. In the union of the two populations allele A, allele B and haplotype AB have frequencies

$$\begin{aligned} p &= \lambda p^1 + (1-\lambda)p^2, \\ q &= \lambda q^1 + (1-\lambda)q^2, \\ h &= \lambda h^1 + (1-\lambda)h^2. \end{aligned}$$

9.1 Lemma. *For numbers $p^1, p^2, q^1, q^2, \lambda$ in $[0,1]$, define p, q and h as in the preceding display. If $h^1 = p^1q^1$ and $h^2 = p^2q^2$, then $h - pq = \lambda(1-\lambda)(p^1-p^2)(q^1-q^2)$.*

Proof. It is immediate from the definitions that

$$h - pq = \lambda h^1 + (1-\lambda)h^2 - (\lambda p^1 + (1-\lambda)p^2)(\lambda q^1 + (1-\lambda)q^2).$$

Here we insert the equilibrium identities $h^1 = p^1q^1$ and $h^2 = p^2q^2$ and obtain the formula $h - pq = \lambda(1-\lambda)(p^1-p^2)(q^1-q^2)$ by elementary algebra. ■

The assumptions $h^1 = p^1q^1$ and $h^2 = p^2q^2$ of the lemma entail lack of association in the subpopulations, and the difference $h - pq$ measures the association in the whole population. The lemma shows that that $h - pq = 0$ if and only if $p^1 = p^2$ or $q^1 = q^2$. Therefore, the joint population can be far from linkage equilibrium if the marginal frequencies are different, even if both subpopulations are in linkage equilibrium. The variable "subpopulation" is said to act as a *confounder* for association.

Rather than to the alleles at two loci, we can apply the lemma also to association between a single (marker) locus and the disease. We define h^1, h^2 and h as the probability that a random individual in the subpopulations or full population carries allele A and is diseased, p^1, p^2, p as the relative frequencies of allele A in the three populations, and q, q^1, q^2 as the prevalence of the disease. The lemma shows that a disease that is unrelated to the allele A in both subpopulations will be associated to the allele in the full population as soon as both the prevalence of the disease and the relative frequency of A are different in the two subpopulations. If the prevalence of the disease is different, then many alleles A may qualify.

The preceding discussion extends to more than two subpopulations. In fact, with the appropriate interpretations this follows from Lemma 2.7.

9.1.1 Testing for Association

To investigate the existence of association between two loci in a population we sample n individuals at random and determine their genotypes at the two loci of interest. Typically we can observe only unordered genotypes. If the two loci possess possible alleles $A_1, \ldots, A_k$ and $B_1, \ldots, B_l$, respectively, then there are $\frac{1}{2}k(k+1)$ unordered genotypes $\{A_i, A_j\}$ for the first locus and $\frac{1}{2}l(l+1)$ unordered genotypes $\{B_u, B_v\}$ for the second locus. Each individual can (in principle) be classified for both loci, yielding data in the form of a $\bigl(\frac{1}{2}k(k+1) \times \frac{1}{2}l(l+1)\bigr)$-table N, with the coordinate N_{ijuv} counting the total number of individuals in the sample with unordered genotypes $\{A_i, A_j\}$ and $\{B_u, B_v\}$ at the two loci. Table 9.1 illustrates this for $k = l = 2$.

	$\{B_1, B_1\}$	$\{B_1, B_2\}$	$\{B_2, B_2\}$
$\{A_1, A_1\}$	N_{1111}	N_{1112}	N_{1122}
$\{A_1, A_2\}$	N_{1211}	N_{1212}	N_{1222}
$\{A_2, A_2\}$	N_{2211}	N_{2212}	N_{2222}

Table 9.1. Two-way classification of a sample for unordered genotypes on two loci with possible alleles A_1, A_2 and B_1, B_2, respectively.

If the n individuals are sampled at random from the population, then the matrix N is multinomially distributed with parameters n and a $\bigl(\frac{1}{2}k(k+1) \times \frac{1}{2}l(l+1)\bigr)$ probability matrix g, which contains the cell relative frequencies of the table in the population. If we do not make assumptions on the structure of g, then its maximum likelihood estimator is the matrix N/n of relative frequencies in the sample. We consider three ways of restricting the matrix g. Write g_{ijuv} for the relative frequency of cell $\{A_i, A_j\} \times \{B_u, B_v\}$.

Under combined Hardy-Weinberg and linkage equilibrium (HW+LE) , the population frequencies g can be expressed in the marginal probabilities (p_i) and (q_u) of the alleles A_i and B_u at the two loci, through the formulas, for every $i \neq j$ and $u \neq v$,

$$\begin{aligned} g_{iiuu} &= p_i^2 q_u^2, \\ g_{iiuv} &= p_i^2 2q_u q_v, \\ g_{ijuu} &= 2p_i p_j q_u^2, \\ g_{ijuv} &= 2p_i p_j 2q_u q_v. \end{aligned} \tag{9.2}$$

The factors 2 arise because the genotypes in the margins of Table 9.1 (and its generalization to loci with more than two alleles) are unordered. The marginal frequencies p_i and q_u constitute $(k-1)+(l-1)$ free parameters, and have as their maximum likelihood estimates the marginal frequencies of the alleles in the sample.

An assumption of random mating (RM) is often considered reasonable, and implies that a genotype is constituted of two randomly chosen haplotypes from the population. If h_{iu} is the frequency of the haplotype A_iB_u in the population, then this assumption implies that, for every $i \neq j$ and $u \neq v$,

$$\begin{aligned} g_{iiuu} &= h_{iu}^2, \\ g_{iiuv} &= 2h_{iu}h_{iv}, \\ g_{ijuu} &= 2h_{iu}h_{ju}, \\ g_{ijuv} &= 2h_{iu}h_{jv} + 2h_{iv}h_{ju}. \end{aligned} \tag{9.3}$$

The sum in the last line arises, because both (unordered) pairs of haplotypes

$$\left\{ \begin{pmatrix} A_i \\ B_u \end{pmatrix}, \begin{pmatrix} A_j \\ B_v \end{pmatrix} \right\}, \qquad \text{and} \qquad \left\{ \begin{pmatrix} A_i \\ B_v \end{pmatrix}, \begin{pmatrix} A_j \\ B_u \end{pmatrix} \right\} \tag{9.4}$$

give rise to the (unordered) genotypes $\{A_i, A_j\}$ and $\{B_u, B_v\}$. This model is parametrized by kl haplotype relative frequencies, which are a set of $kl-1$ free parameters.

Third it is possible to assume that the genotypes of the two loci are independent (LE) without assuming random mating. This assumption can be described directly in terms of the observed unordered genotypes at the two loci, and comes down to the assumption of independence of the two margins in Table 9.1 and its generalization to higher-dimensional tables (see Section 14.1.5). The parameters of this model are the marginal relative frequencies of the unordered genotypes at the two loci, of which there are $\frac{1}{2}k(k+1) - 1 + \frac{1}{2}l(l+1) - 1$.

Thus we have a model of dimension $\frac{1}{2}k(k+1)\frac{1}{2}l(l+1) - 1$ leaving the matrix g free, the model RM parametrized by the $kl-1$ haplotype frequencies, the model HW+LE of dimension $k-1+l-1$ parametrized by the marginal allele frequencies, and the model LE of dimension $\frac{1}{2}k(k+1) + \frac{1}{2}l(l+1) - 2$ parametrized by the unordered genotype frequencies at the two loci. The models are nested and the smaller models can be tested in their containing models by the likelihood ratio or chisquare statistic.

(i) The validity of the submodel HW+LE can be tested within the full model on $\frac{1}{2}k(k+1)\frac{1}{2}l(l+1)-1-(k-1)-(l-1)$ degrees of freedom.
(ii) The assumption (9.3) of random mating RM can be tested within the full model on $\frac{1}{2}k(k+1)\frac{1}{2}l(l+1)-kl$ degrees of freedom.
(iii) The model HW+LE can be tested within the model RM on $kl-1-(k-1+l-1)=(k-1)(l-1)$ degrees of freedom.
(iv) The model LE can be tested within the full model on $\bigl(\frac{1}{2}k(k+1)-1\bigr)\bigl(\frac{1}{2}l(l+1)-1\bigr)$ degrees of freedom.

This is all under the assumption that under the null hypothesis none of the frequencies are zero (to protect the level of the test), and with the understanding that the test has power mostly against alternatives that deviate from the null in a significant number of cells (as it "spreads" its sensitivity over all cells of the table). Within the present context the third and fourth tests are the most interesting ones. They both test for the absence of association (9.2) of the two loci, where the third test assumes random mating (and will be preferable if that is a correct assumption) and the fourth has the benefit of being very straightforward.

9.1.2 Estimating Haplotype Frequencies

The maximum likelihood estimator of the haplotype frequencies (h_{iu}) under the random mating model maximizes the likelihood

$$(h_{iu})\mapsto\binom{n}{N}\prod_{i,u}h_{iu}^{2N_{iiuu}}\prod_{i,u\neq v}(2h_{iu}h_{iv})^{N_{iiuv}}\prod_{i\neq j,u}(2h_{iu}h_{ju})^{N_{ijuu}}$$
$$\times\prod_{i\neq j,u\neq v}(2h_{iu}h_{jv}+2h_{iv}h_{ju})^{N_{ijuv}}.$$

Because direct computation of the point of maximum is not trivial, it is helpful to carry out the maximization using the EM-algorithm. We take the "full data" equal to the frequencies of the unordered pairs of haplotypes, and the observed data the matrix N, as exemplified for $k=l=2$ in Table 9.1. For individuals who are homozygous at one of the two loci (or both loci) the full data is observable. For instance, an individual classified as $\{A_i,A_i\}$ and $\{B_u,B_v\}$ clearly possesses the unordered pair of haplotypes

$$\Bigl\{\begin{pmatrix}A_i\\B_u\end{pmatrix},\begin{pmatrix}A_i\\B_v\end{pmatrix}\Bigr\}. \tag{9.5}$$

On the other hand, the haplotypes of individuals that are heterozygous at both loci cannot be resolved from the data. The EM-algorithm can be understood as splitting the observed numbers N_{ijuv} of pairs of genotypes $\{A_i,A_j\}$ and $\{B_u,B_v\}$ with $i\neq j$

and $u \neq v$ recursively into $\hat{N}_{iu,jv|1}$ and $\hat{N}_{iv,ju|1}$ pairs of haplotypes as given in (9.4) by the formulas

$$
\begin{aligned}
\hat{N}_{iu,jv|1} &= N_{ijuv}\frac{h_{iu|0}h_{jv|0}}{h_{iu|0}h_{jv|0}+h_{iv|0}h_{ju|0}},\\
\hat{N}_{iv,ju|1} &= N_{ijuv}\frac{h_{iv|0}h_{ju|0}}{h_{iu|0}h_{jv|0}+h_{iv|0}h_{ju|0}}.
\end{aligned}
\tag{9.6}
$$

Here the $h_{iu|0}$ are the current iterates of the EM-algorithm. Given these reconstructions of the haplotypes, the EM-algorithm computes new haplotype estimates $h_{iu|1}$ from the empirical haplotype frequencies, and proceeds to the next iteration.

9.7 EXERCISE. For two biallelic loci there are 9 combinations of unordered genotypes, as given in Table 9.1. There are 4 different haplotypes and 10 different combinations of two unordered haplotypes. Show that 8 of the cells of Table 9.1 uniquely define a pair of unordered haplotype and one cell corresponds to two of such pairs. Which cell?

To consider this in more detail define $Y_{iu,jv}$ to be the number of individuals in the sample with unordered pair of haplotypes (9.5). We may think of the vector $(Y_{iu,jv})$ as arising from counting the number of pairs of haplotypes after first generating $2n$ haplotypes, haplotype A_iB_u appearing with probability h_{iu}, next forming the n pairs consisting of the first and second, the third and fourth, and so on. The likelihood for observing the vector $(Y_{iu,jv})$ is proportional to

$$
\prod_{i\neq j \text{ or } u\neq v}(2h_{iu}h_{jv})^{Y_{iu,jv}}\prod_{i,u}(h_{iu}^2)^{Y_{iu,iu}} \propto \prod_{i,u} h_{iu}^{N_{iu}},
\tag{9.8}
$$

where N_{iu} is the total number of haplotypes A_iB_u in the sample of $2n$ haplotypes. It follows from this that the maximum likelihood estimator for (h_{iu}) based on observing $(Y_{iu,jv})$ is equal to the maximum likelihood estimator based on (N_{iu}), which is simply the vector of relative frequencies $(N_{iu}/2n)$.

In fact we observe only the numbers of N_{ijuv} of pairs of unordered genotypes. The preceding discussion shows that

$$
N_{ijuv} = \begin{cases} Y_{iu,jv} & \text{if } i=j \text{ or } v=u,\\ Y_{iu,jv}+Y_{iv,ju} & \text{if } i\neq j \text{ and } v\neq u.\end{cases}
$$

The E-step of the EM-algorithm computes the conditional expectation given $N = (N_{ijuv})$ of the logarithm of the likelihood (9.8),

$$
\mathrm{E}_0\Bigl(\sum_{i\neq j \text{ or } u\neq v} Y_{iu,jv}\log(2h_{iu}h_{jv}) + \sum_{i,u} Y_{iu,iu}\log(h_{iu}^2)|\, N\Bigr),
$$

where the subscript 0 on the expectation indicates to use the current value of the haplotype frequencies. By the form of the likelihood the computation comes down to computing the conditional expectations $\mathrm{E}_0(Y_{iu,jv}|\, N)$ for every coordinate (iu, jv).

For the haplotypes that can be uniquely resolved from the genotypes ($i = j$ or $u = v$) this conditional expectation is simply $\mathrm{E}_0(Y_{iu,jv}|\, N) = N_{ijuv}$. For the other haplotypes ($i \neq j$ and $u \neq v$) the N_{ijuv} pairs must be partitioned into $Y_{iu,jv} + Y_{iv,ju}$. Because the two pairs of haplotypes (9.4) occur in the population with probabilities $2h_{iu}h_{jv}$ and $2h_{iv}h_{ju}$, respectively, we obtain that $\mathrm{E}_0(Y_{iu,jv}|\, N)$ is given by $\hat{N}_{iu,jv|1}$ as given previously.

After thus replacing the unobserved frequencies $Y_{iu,jv}$ by their conditional expectations, in the M-step we maximize the likelihood with respect to (h_{iu}). As noted this leads to the empirical estimates based on the total numbers of haplotypes of the various types among the $2n$ haplotypes.

9.1.3 Measures of Linkage Disequilibrium

An obvious quantitative measure of linkage disequilibrium between loci with alleles A_i and B_j with haplotype frequencies (h_{ij}) and marginal frequencies (p_i) and (q_j) is

$$D_{ij} = h_{ij} - p_i q_j. \tag{9.9}$$

These quantities are the difference between the "joint" probability of the alleles at the two loci (the probabilities of the haplotypes A_iB_j) and the probabilities if the loci were independent. The measure is illustrated for two biallelic loci in Table 9.2. In principle there are four measures D_{ij} for this table, but these can be summarized by just one of them.

9.10 Lemma. *For two biallelic loci* $D_{11} = D_{22} = -D_{12} = -D_{21}$.

Proof. Because $h_{22} = 1 - p_1 - q_1 + h_{11}$ it follows that $D_{22} = 1 - p_1 - q_1 + h_{11} - (1 - p_1)(1 - q_1) = D_{11}$. Similarly, because $h_{12} = 1 - h_{11} - q_2$ it follows that $D_{12} = 1 - h_{11} - (1 - q_1) - p_1(1 - q_1) = -D_{11}$. The relationship $D_{12} = D_{21}$ follows by symmetry (exchange 1 and 2 for one of the loci) or by similar reasoning. ∎

	B_1	B_2	
A_1	h_{11}	h_{12}	p_1
A_2	h_{21}	h_{22}	p_2
	q_1	q_2	1

Table 9.2. Table of haplotype frequencies for two-loci haplotypes, one with with alleles A_1 and A_2 and the other with alleles B_1 and B_2.

The range of the numbers D_{ij} is restricted through the marginal allele frequencies. This is shown by the inequalities in the following lemma. Such restrictions seem undesirable for a measure of dependence. For instance, if one of the alleles A_i or B_j is rare or very abundant, then $D_{i,j}$ is automatically close to zero, even though the marginal frequency is not informative on the joint distribution.

9.11 Lemma. $-\min\bigl(p_iq_j,(1-p_i)(1-q_j)\bigr)\le D_{ij}\le \min\bigl(p_i(1-q_j),(1-p_i)q_j\bigr)$.

Proof. Consider without loss of generality the alleles A_1 and B_1. Because we can lump together the alleles $A_2,\ldots,A_k$ and $B_2,\ldots,B_l$ into a single allele, we can assume also without loss of generality that the loci are biallelic.

The inequalities $0\le h_{11}\le p_1$ and the definition of D_{11} immediately imply that $-p_1q_1\le D_{11}\le p_1(1-q_1)$, which are two of the four inequalities given by the lemma. From an application of these inequalities to D_{22} we obtain by symmetry that $-p_2q_2\le D_{22}\le p_2(1-q_2)$. Because $D_{11}=D_{22}$ this gives the remaining two inequalities of the lemma. ∎

In view of the preceding lemma the numbers

$$D'_{ij}=\begin{cases}\dfrac{D_{ij}}{p_iq_j\wedge(1-p_i)(1-q_j)}, & \text{if } D_{ij}\le 0,\\[2ex] \dfrac{D_{ij}}{p_i(1-q_j)\wedge(1-p_i)q_j}, & \text{if } D_{ij}\ge 0,\end{cases}$$

are contained in the interval $[-1,1]$. Inspection of the proof of the preceding lemma reveals that the extremes -1 and 1 are attained if a diagonal element in Table 9.2 is zero ($D'_{ij}=-1$) or an off-diagonal element is zero ($D'_{ij}=1$).

An alternative standardization of D_{ij} is

$$r_{ij}=\Delta_{ij}=\frac{D_{ij}}{\sqrt{p_i(1-p_i)q_j(1-q_j)}}. \tag{9.12}$$

This is the correlation coefficient between two indicator variables 1_{A_i} and 1_{B_j} corresponding to partitions $\cup_iA_i=\cup_jB_j$ of some probability space with $P(A_i\cap B_j)=h_{ij}$. (The probability space can correspond to choosing an haplotype at random and saying that events A_i and B_j occur if the haplotype is A_iB_j.)

The classification of a random sample of n haplotypes for two biallelic loci, as in Table 9.3, divided by n gives a sample version of Table 9.2. The chisquare statistic for independence in Table 9.3 can be written in the form

$$\sum_{i=1}^2\sum_{j=1}^2 n\frac{\bigl(N_{ij}/n-(N_{i\cdot}/n)(N_{\cdot j}/n)\bigr)^2}{(N_{i\cdot}/n)(N_{\cdot j}/n)}=n\frac{\bigl(N_{11}/n-(N_{1\cdot}/n)(N_{\cdot 1}/n)\bigr)^2}{(N_{1\cdot}/n)(N_{2\cdot}/n)(N_{\cdot 1}/n)(N_{\cdot 2}/n)}.$$

The equality is easily derived after noting the fact that the numerators of the four terms in the double sum are in terms of empirical versions $\hat D_{ij}$ of the desequilibrium measures D_{ij}, which have equal absolute versions, by Lemma 9.10. The right side is the empirical version of the measure nr_{ij}^2, which is thus exhibited to be a testing statistic for independence. The usefulness of this observation is limited by the fact that we usually do not completely observe the numbers of haplotypes in Table 9.3.

	B_1	B_2	
A_1	N_{11}	N_{12}	$N_{1.}$
A_2	N_{21}	N_{22}	$N_{2.}$
	$N_{.1}$	$N_{.2}$	n

Table 9.3. Table of haplotype frequencies for two-loci haplotypes each with two alleles: N_{ij} out of n haplotypes are A_jB_j.

9.13 EXERCISE. Show that the inequalities in Lemma 9.11 are sharp. [Hint: as shown in the proof, in the biallelic case the four inequalities are attained if the haplotype frequency in the appropriate cell of the four cells in Table 9.2 is zero.]

9.14 EXERCISE. Show that the bounds on D_{ij} given in Lemma 9.11 are tighter than the bounds that follow from the fact that $|\Delta_{ij}| \leq 1$ (which follows because Δ_{ij} is a correlation coefficient).

9.2 Case-Control Tests

Suppose we measure the genotypes at marker loci for random samples of affected individuals (*cases*) and healthy individuals (*controls*). A *case-control test* is simply a two-sample test for the null hypothesis that the markers of cases and controls the same.

The simplest approach is to consider one marker at a time. Typically we observe for each individual only the unordered genotypes, without phase information. If the marker has alleles $M_1, \ldots, M_l$, then the observations can be summarized by vectors of length $l(l+1)/2$ giving the counts of the genotypes $\{M_i, M_j\}$ in the samples of cases and controls. Under the random sampling assumption, these vectors are independent and multinomially distributed with parameters (n^A, g^A) and (n^U, g^U), respectively, for g^A and g^U the vectors giving the probabilities that a case or control has genotype $\{M_i, M_j\}$. We perform a test of the null hypothesis that the vectors g^A and g^U are equal.

Because many affections are multigenic, it may be fruitful to investigate the joint (indirect) effects of markers. Then we combine the data in two higher-dimensional tables, giving a cross classification of the cases and controls on the various markers. If we observe only unordered genotypes for the individual marker loci, then we do not observe haplotypes, and base the test on sets of unordered genotypes. Even for a small set of markers the tables may have many cells, and testing equality of the probabilities for the tables of cases and controls may be practically impossible without modelling these probabilities through a lower-dimensional parameter. Logistic regression (see Section 9.2.3) is often used for this purpose. Another complication is the very large number of tables that could be formed by selecting all subsets of a given number of marker loci. This creates both a practical computational problem and the theoretical problem of how to correct for multiple testing. Searching for a suitable set of markers is perhaps best viewed as a problem of statistical model selection (See Section 9.2.8).

As noted in the introduction of the chapter, linkage disequilibrium between causal and marker loci is necessary for these approaches to have chance of success. To gain quantitative insight in this, assume that the disease is caused by the genes at k loci, and denote the possible haplotypes at these loci by $D_1, \ldots, D_R$. If we could observe the haplotypes at the causal loci and directly compare the frequencies of the genotypes (D_r, D_s) among cases and controls, then we would be comparing two multinomial vectors with vectors of success probabilities $P(D_{r,s}|\, A)$ and $P(D_{r,s}|\, U)$, respectively, for $D_{r,s}$ the event that an individual has genotype (D_r, D_s), and A and U the events that an individual is affected (i.e. a case) or unaffected (i.e. a control). By Bayes' rule these probability vectors can be expressed in the penetrance $f_{r,s} = P(A|\, D_{r,s})$ of the disease as

$$\begin{aligned} P(D_{r,s}|\, A) &= \frac{f_{r,s} p_{r,s}}{P(A)}, \\ P(D_{r,s}|\, U) &= \frac{(1 - f_{r,s}) p_{r,s}}{P(U)}. \end{aligned} \tag{9.15}$$

Here $p_{r,s}$ is the relative frequency of genotype (D_r, D_s) in the population, and $P(A)$ is the *prevalence* of the disease, which can be expressed in the penetrance and haplotype frequencies as $P(A) = \sum_r \sum_s f_{r,s} p_{r,s}$. The power to detect the causal locus depends on the magnitude of the difference of these probabilities.

9.16 EXERCISE. Formula (9.15) suggests that the power of a case-control test depends on the prevalence of the disease. Given that in practice cases and controls are sampled separately and independently, is this surprising? Explain. [Hint: consider (9.15) in the special case of full penetrance without phenocopies, i.e. $f_{r,s} \in \{0, 1\}$ for every (r, s).]

In reality we may base the case-control test on a marker locus that is not causal for the affection. If we use a marker with possible haplotypes $M_1, \ldots, M_l$, then the relevant probabilities are not the ones given in the preceding display, but the probabilities $P(M_{i,j}|\, A)$ and $P(M_{i,j}|\, U)$, for $M_{i,j}$ the event that an individual has marker genotype (M_i, M_j). We interprete the notion of "causal locus" (as usual) to mean that marker genotype and affection status are conditionally independent given the causal genotype, i.e. $P(M_{i,j}|\, D_{r,s} \cap A) = P(M_{i,j}|\, D_{r,s})$. Then the marker probabilities can be written in the form

$$\begin{aligned} P(M_{i,j}|\, A) &= \sum_r \sum_s P(M_{i,j}|\, D_{r,s}) P(D_{r,s}|\, A) = \sum_r \sum_s \frac{h_{ir,js}}{p_{r,s}} P(D_{r,s}|\, A), \\ P(M_{i,j}|\, U) &= \sum_r \sum_s P(M_{i,j}|\, D_{r,s}) P(D_{r,s}|\, U) = \sum_r \sum_s \frac{h_{ir,js}}{p_{r,s}} P(D_{r,s}|\, U). \end{aligned} \tag{9.17}$$

Here $h_{ir,js}$ is the relative frequency of the genotype (M_iD_r, M_jD_s), for M_iD_r the haplotype formed by uniting the the marker haplotype M_i with disease haplotype D_r.

event	*interpretation*	*probability*
A	individual is affected	$P(A)$
U	individual is unaffected	$P(U)$
$M_{i,j}$	individual has marker genotype (M_i, M_j)	$q_{i,j}$
$D_{r,s}$	individual has disease genotype (D_r, D_s)	$p_{r,s}$
M_i	individual has paternal marker haplotype M_i	q_i
D_r	individual has paternal disease haplotype D_r	p_r

Table 9.4. Events.

If the disease and marker loci are not associated, then $P(M_{i,j}|\, D_{r,s}) = P(M_{i,j})$ for all r, s, i, j, and both probabilities can be seen to reduce to the unconditional probability $P(M_{i,j})$. In the other case, we may hope that a difference between the probability vectors (9.15) is translated into a difference between the corresponding marker probabilities (9.17). That the latter are mixtures of the first suggests that the difference is attenuated by going from the causal probabilities (9.15) to the marker probabilities (9.17). The hope is that this attenuation is small if the marker and causal loci are close, and increases as the marker loci move away from the causal loci, so that the null hypothesis of no difference between case and control marker probabilities is rejected if, and only if, a marker locus is close to a causal locus.

The calculation reveals that the marker probabilities will be different as soon as the conditional probabilities $P(M_{i,j}|\, D_{r,s})$ possess a certain pattern. The location of the marker loci relative to the causal loci is important for this pattern, but so may be other variables. Spurious association between marker and causal loci, for instance due to population structure as discussed in Section 9.1, makes the probabilities $P(M_{i,j}|\, D_{r,s})$ differ from the unconditional probabilities $P(M_{i,j})$, and may create a similar pattern. In that case the null hypothesis of no difference between the marker probabilities (9.17) may be correctly rejected without the marker being close to the causal loci. This is an important drawback of association studies: proper control of confounding variables may be necessary.

If the marker loci under investigation happen to be associated to only a subset of causal loci, in the sense that the probabilities $P(M_{i,j}|\, D_{r,s})$ are equal to $P(M_{i,j}|\, \bar D_{\bar r,\bar s})$ for $\bar D_{\bar r,\bar s}$ referring to the pair of haplotypes at a subset of causal loci (i.e. a union of events $D_{r,s}$ over the subset of (r, s) with $\bar r \subset r$ and $\bar s \subset s$ fixed), then (9.17) is valid with $D_{r,s}$ replaced by $\bar D_{\bar r,\bar s}$. If we investigate a single marker locus, then it seems not impossible that only a small set of disease loci is associated. Note that we are referring to association, i.e. dependence in the population, and not to linkage. For instance, it cannot be a-priori excluded that a marker locus on one chromosome is associated to a disease locus on another chromosome.

Summing over j in (9.17) yields the probabilities $P(M_i|\, A)$ an $P(M_i|\, U)$ of a diseased or healthy person having paternal marker haplotype M_i. This is mostly of interest if the population is in Hardy-Weinberg equilibrium (at the haplotype level), so that the distribution of its genotypes can be expressed in the haplotype relative frequencies. Under the assumption of Hardy-Weinberg,

$$\begin{aligned} P(M_i|\,A) &= \sum_r\sum_s P(M_i|\,D_{r,s})P(D_{r,s}|\,A) = \sum_r \frac{h_{ir}}{p_r}P(D_r|\,A),\\ P(M_i|\,U) &= \sum_r\sum_s P(M_i|\,D_{r,s})P(D_{r,s}|\,U) = \sum_r \frac{h_{ir}}{p_r}P(D_r|\,U).\end{aligned} \tag{9.18}$$

9.2.1 Chisquare Tests

The family of chisquare tests is a standard choice for testing hypotheses on multinomial vectors. In the present situation the cases and controls generate two independent multinomial tables, and the null hypothesis is that these tables have equal cell probabilities. The tables and the chisquare test can be set up in various ways.

For a test based on a single marker we form the multinomials tables as the counts of unordered marker genotypes $\{M_i, M_j\}$. The probability vectors of these tables could be left completely unspecified, and we could perform the standard chisquare test for comparing two multinomial tables, discussed in Section 14.1.6. For a single marker with l alleles, this yields a chisquare test on $l(l+1)/2-1$ degrees of freedom. An advantage of this approach is that we do not need to make assumptions, such as Hardy-Weinberg equilibrium. A disadvantage may be that the number of cells of the multinomial table may be large, and the test may have poor power and/or the null distribution of the test statistic may be badly approximated by the chisquare distribution. A well known rule of thumb to protect the level is that under the null hypothesis the expected frequency of each cell in the table must be at least five. However, this rule of thumb does not save the power, the problem being that the chisquare test distributes its power over all possible deviations of the cell probabilities from the null hypothesis. This is a reasonable strategy if the deviation from the null hypothesis consists indeed of small deviations in many cells, but bigger deviations in only a few cells may go undetected. For a multiallelic marker locus it is not improbable that only a single allelic value is linked to the disease.

An alternative is to model the cell frequencies through a smaller number of parameters. One possibility is to use the forms for $P(M_{i,j}|\,A)$ and $P(M_{i,j}|\,U)$ found in (9.15)-(9.17), but since the penetrance parameters $f_{r,s}$ and/or genotype frequencies are typically not known, this will not reduce the dimension. If the population is thought to be in Hardy-Weinberg or linkage equilibrium, at least under the null hypothesis, then it makes sense to restrict the cell probabilities to observe the implied relations. Structural models in terms of main and interaction effects can also be used within the chisquare context, but are usually implemented within the context of logistic regression, discussed in Section 9.2.3.

If the population is in Hardy-Weinberg equilibrium at the marker locus, at least under the null hypothesis, then each individual case or control contributes two independent marker alleles. Then the total counts of alleles in cases and controls are multinomially distributed with sample sizes twice the numbers of cases and controls. Performing a chisquare test on these haplotype counts (and not the genotypes) reduces dimension, and gives a more powerful test if the assumption of Hardy-Weinberg equilibrium is valid. On the other hand, in disequilibrium the total allele counts would not be multinomial, and, failing a specific alternative model for Hardy-Weinberg equilibrium, it is better to use the genotype counts.

For tests based on multiple markers we can follow the same strategies. The simplest approach is to form a multinomial table with as cells the possible combinations of unordered genotypes across the loci. We might leave the corresponding cell probabilities free, or could restrict them to satisfy various equilibrium assumptions.

Power. The (local) power of the chisquare test for comparing two independent multinomial tables can be expressed in a *noncentrality parameter* (see Section 14.1.6). If the multinomial tables have m^A and m^U replicates and probability vectors q^A and q^U, then the square noncentrality parameter for testing $H_0: q^A = q^U$ is equal to

$$\frac{m^A m^U}{m^A + m^U}\left\|\frac{q^A - q^U}{\sqrt{q}}\right\|^2, \tag{9.19}$$

where q is the common value of the probabilities q^A and q^U under the null hypothesis. (This noncentrality parameter refers to the “local” power in the sense of power for alternatives close to the null hypothesis of no difference between cases and controls, i.e. $q^A \approx q^U$. It has nothing to do with proximity on the genome.)

In the present situation, if the test were based on the ordered genotypes, the vectors q^A and q^U would be taken equal to the vectors with coordinates

$$q^A_{i,j} = P(M_{i,j}|\, A), \qquad q^U_{i,j} = P(M_{i,j}|\, U).$$

The null probability q would be the vector of marker genotype relative frequencies, with coordinates $q_{i,j} = P(M_{i,j})$. In the more realistic situation of unobserved phases these probabilities would be replaced by the corresponding probabilities of unordered marker genotypes. Alternatively, under Hardy-Weinberg equilibrium a single marker test would be based on the total allele counts, and the relevant probabilities are the vectors with coordinates

$$q^A_i = P(M_i|\, A), \qquad q^U_i = P(M_i|\, U).$$

In this case the null probabilities are the marker haplotype frequencies q_i, and the sample sizes m^A and m^U would be twice the numbers of cases and controls. We abuse notation by using the same symbol for genotypic and haplotypic relative frequencies. The context or the subscript (double or single) will make clear which of the two is involved.

In (9.17) and (9.18) we have seen that the probability vectors q^A and q^U can differ only if the marker loci are associated to a disease locus. It can be expected that the difference is larger if the association is stronger. It is instructive to make this precise by comparing the power of the test based on the markers to the power that would have been obtained had the test been based on the causal loci. In the latter case, for the test based on genotypes, the relevant probability vectors would be p^A and p^U with coordinates

$$p^A_{r,s} = P(D_{r,s}|\, A), \qquad p^U_{r,s} = P(D_{r,s}|\, U).$$

For the test based on haplotypes the relevant probability vectors p^A and p^U would be given by (again with abuse of notation)

$$p^A_r = P(D_r|\, A), \qquad p^U_r = P(D_r|\, U).$$

Equations (9.17) and (9.18) give the relationships between the marker probabilities and causal probabilities. They can be written in the matrix forms

$$\Bigl(\frac{q^A_{i,j} - q^U_{i,j}}{\sqrt{q}_{i,j}}\Bigr) = \Bigl(\frac{h_{ir,js}}{\sqrt{q_{i,j}}\sqrt{p_{r,s}}}\Bigr)\Bigl(\frac{p^A_{r,s} - p^U_{r,s}}{\sqrt{p_{r,s}}}\Bigr), \tag{9.20}$$

for the test based on genotypes, and for the haplotypic test

$$\Bigl(\frac{q^A_i - q^U_i}{\sqrt{q}_i}\Bigr) = \Bigl(\frac{h_{ir}}{\sqrt{q_i}\sqrt{p_r}}\Bigr)\Bigl(\frac{p^A_r - p^U_r}{\sqrt{p}_r}\Bigr). \tag{9.21}$$

The square noncentrality parameters of the tests based on the marker and causal loci are proportional to the square norms of the vectors on the left and the far right. As shown in (9.19) the proportional factors are $n^2/(n+n) = n/2$ and $(2n)^2/(2n+2n) = n$ if the numbers of cases and controls are both equal to n.[♯] This shows that the asymptotic relative efficiency of the tests is given by the square noncentrality parameter. As expected using marker loci is always less efficient.

9.22 Lemma. *The vectors* $(q^A - q^U)/\sqrt{q}$ *and* $(p^A - p^U)/\sqrt{p}$ *in (9.20) and in (9.21) satisfy*

$$\bigl\|(q^A - q^U)/\sqrt{q}\bigr\|^2 \le \bigl\|(p^A - p^U)/\sqrt{p}\bigr\|^2.$$

Proof. We prove the lemma for the haplotypic case (9.21). The proof for the genotype-based comparison is similar. For notational convenience let $\Omega = \cup_i M_i = \cup_r D_r$ be two partitions of a given probability space such that $P(M_i \cap D_r) = h_{ir}$ for every (i,r), and define a random vector U from the first partition by $U = \bigl((1_{M_1} - p_1)/\sqrt{p_1}, \ldots, (1_{M_k} - p_k)/\sqrt{p_k}\bigr)^T$ and similarly define a random vector V from the second partition.

Because $(p^A - p^U)^T 1 = 0$, the matrix $\bigl(h_{ir}/(\sqrt{q_i}\sqrt{p_r})\bigr)$ in (9.21) can be replaced by the matrix $\bigl((h_{ir} - q_i p_r)/(\sqrt{q_i}\sqrt{p_r})\bigr) = \mathrm{E}UV^T$. We wish to show that this matrix has norm smaller than 1. By the Cauchy-Schwarz inequality $(x^T\mathrm{E}UV^Ty)^2 = \bigl(\mathrm{E}(x^TU)(V^Ty)\bigr)^2 \le \mathrm{E}(x^TU)^2\mathrm{E}(y^TV)^2$. Now $\mathrm{E}(x^TU)^2 =$

[♯] The *relative efficiency* of two family of tests for testing the same simple null hypothesis against the same simple alternative hypothesis is defined as the quotient n/n' of the numbers of observations needed with the two tests to achieve a given level α and power $1-\beta$. In general this depends on the two hypotheses and on the pair (α, β), but often it does not.

$x^T \mathrm{E} UU^T x$, where $\mathrm{E} UU^T = I - \sqrt{q}\sqrt{q}^T$ is a projection matrix and hence $\mathrm{E}(x^T U)^2 \leq \|x\|^2$. The mean $\mathrm{E}(y^T V)^2$ satisfies the analogous inequality. We conclude that $x^T \mathrm{E} UV^T y \leq \|x\| \|y\|$, and hence the norm of the matrix $\mathrm{E} UV^T$ is bounded by 1. ∎

9.23 Example (Biallelic marker and disease locus). The relative efficiency of comparing marker and causal loci takes a simple form in the case both loci are biallelic and the test used is based on the allele counts (under the assumption of Hardy-Weinberg equilibrium). In this situation the multinomial vectors have only two classes, and the tests compare two binomial probabilities. As shown in Example 14.12 the noncentrality parameter (9.19) can be written as $2n^A 2n^U/(2n^A + 2n^U)(q_1^A - q_1^U)^2/(q_1 q_2)$. Equation (9.21) shows that

$$\frac{q_1^A - q_1^U}{\sqrt{q_1}} = \frac{h_{11}}{\sqrt{q_1}\sqrt{p_1}} \frac{p_1^A - p_1^U}{\sqrt{p_1}} + \frac{h_{12}}{\sqrt{q_1}\sqrt{p_2}} \frac{p_2^A - p_2^U}{\sqrt{p_2}} = \frac{p_1^A - p_1^U}{\sqrt{q_1}} \Bigl(\frac{h_{11}}{p_1} - \frac{h_{12}}{p_2}\Bigr).$$

The asymptotic relative efficiency of the tests on comparing the total count of marker alleles versus the test based on comparing the causal alleles is therefore given by the quotient

$$\frac{(q_1^A - q_1^U)^2/(q_1 q_2)}{(p_1^A - p_1^U)^2/(p_1 p_2)} = \Bigl(\frac{h_{11}}{p_1} - \frac{h_{12}}{p_2}\Bigr)^2 \frac{p_1 p_2}{q_1 q_2} = r_{11}^2,$$

where r_{11} is the measure of linkage disequilibrium between disease and marker locus given in (9.12).

Being a correlation, the relative efficiency is smaller than 1: using the marker instead of the causal locus requires $1/r_{11}^2$ as many observations to achieve the same power. This calculation appears to be the basis of the believe that in genome-wide association studies it is sufficient to include enough markers so that any locus has correlation higher than some cut-off (e.g. 0.8) with one of the marker loci. □

Warning. The test based on allele count assumes Hardy-Weinberg equilibrium. Is this a reasonable assumption also under the alternative that there is difference between cases and controls? Or should one correct the power for dependence? Maybe not the local power under the assumption of Hardy-Weinberg under the null hypothesis?

9.24 Example (Additive penetrance). □

* 9.2.2 Fisher's Exact Test

The chisquare tests for multinomial tables derive their name from the approximation of the distribution of the test statistic under the null hypothesis, valid for large sample sizes. Among the tests that are not based on approximations, *Fisher's exact test* for the 2 × 2-table (comparing two binomial distributions) is the best known. The extra effort to implement it may be worth while for not too large sample sizes.

* 9.2.3 Logistic Regression

The logistic regression model gives the flexibility of modelling the effects of several loci together in various ways, and is particularly attractive if the number of alleles or loci is large. It also permits incorporation of additional background variables into the analysis, for instance covariates such as sex or age, but also variables meant to control for population structure or other confounding factors.

Although our intended application is the case-control setting, where the numbers of cases and controls are fixed in advance, the logistic regression is easiest to describe in the set-up of a random sample from the full population. The data on one individual then consists of an indicator Y for case-control status ($Y = 1$ for a case and $Y = 0$ for a control), and information X on the genotype of the individual and possible covariates. If X is coded as a vector with values in $\mathbb{R}^k$, then the logistic regression model postulates that, for some vector $\beta \in \mathbb{R}^k$,

$$P(Y = 1|\, X) = \frac{1}{1 + e^{-\beta^T X}}.$$

An equivalent formulation of this model is that the *log odds ratio* satisfies the linear relationship

$$\log \frac{P(Y = 1|\, X)}{P(Y = 0|\, X)} = \beta^T X.$$

We investigate the importance of a coordinate X_j of X by testing whether the jth coordinate β_j of β is zero.

Both the likelihood ratio and the score test are standard for this purpose, with the second being preferable if computational simplicity counts, as in a genome-wide analysis, where the test is applied many times, for various (sets of) marker loci. For $\Psi(x) = 1/(1 + e^{-x})$ the logistic function, the log likelihood for one individual can be written

$$Y \log \Psi(\beta^T X) + (1 - Y) \log\bigl(1 - \Psi(\beta^T X)\bigr) = Y\beta^T X - \log(1 + e^{\beta^T X}).$$

Noting that $\Psi' = \Psi(1 - \Psi)$, we can compute the score-function and Fisher information matrix as

$$\begin{aligned} \dot\ell_\beta(Y|\, X) &= \bigl(Y - \Psi(\beta^T X)\bigr)X, \\ I_\beta = -\mathrm{E}_\beta \ddot\ell_\beta(Y|\, X) &= \mathrm{E}_\beta \Psi'(\beta^T X) X X^T. \end{aligned}$$

We assume that the Fisher information matrix is nonsingular. In the case-control setting the expectation must be understood relative to a vector X equal to the independent variable of an individual chosen randomly from the collection of cases and controls weighted by the fractions cases and controls in the total sample.

If $\hat\beta_0$ is the maximum likelihood estimator of β under the null hypothesis and $\hat p_{i,0} = \Psi(\hat\beta_0^T X^i)$, then the score test statistic takes the form

$$(9.25)\qquad \sum_{i=1}^n (Y^i - \hat p_{i,0})(X^i)^T \Bigl(\sum_{i=1}^n \hat p_{i,0}(1 - \hat p_{i,0}) X^i (X^i)^T\Bigr)^{-1} \sum_{i=1}^n (X^i)^T (Y^i - \hat p_{i,0}).$$

Under the conditions that the null hypothesis is correctly specified and contains the true parameter as a relative inner point this statistic possesses approximately a chisquare distribution, with degrees of freedom equal to k minus the (local) dimension of the null hypothesis. If the observations were sampled independently from a population, then this follows from Theorem 14.33. A proof for the case-control setting can follow the same lines, the essence being that the score statistic $n^{-1/2}\sum_{i=1}^n \dot\ell_\beta(Y^i|\,X^i)$ is asymptotically Gaussian and the average $n^{-1}\sum_{i=1}^n \Psi'(\beta^T X)XX^T$ tends to the Fisher information matrix.

The sampling of the individuals according to the case-control design is the preferred choice in practice, as it will increase the power of the test. This is true in particular if the number of cases in the population is small, so that a random sample from the population would typically consist mostly of controls. However, as indicated in the analysis the difference between the two designs can be ignored. A closer inspection (see Section 14.5) reveals that the prevalence of the affection in the population is, of course, not estimable from the case-control design, but the coefficients β_j of nontrivial covariates X_j are identifiable, and the (profile) likelihood functions for the two models are proportional for these coefficients.

9.26 Example (Full null hypothesis). Suppose that the independent vector $X = (1, X_1, \ldots, X_k)$ contains an intercept as its first coordinate, and the null hypothesis is that the coefficients $\beta_1, \ldots, \beta_k$ of the other coordinates $X_1, \ldots, X_k$ are zero. Under the null hypothesis the probability $P(Y = 1|\,X) = \Psi(\beta_0)$ does not depend on X, and is a free parameter. Therefore, the maximum likelihood estimator of $\beta = (\beta_0, \ldots, \beta_k)$ under the null hypothesis is $\hat\beta_0 = (\hat\beta_{00}, 0, \ldots, 0)$ for $\Psi(\hat\beta_{00})$ the maximum likelihood estimator of a binomial proportion: $\Psi(\hat\beta_{00}) = \bar Y$, for $\bar Y$ the the proportion of cases, and hence $\bar\beta_{00} = \log\bigl(\bar Y/(1-\bar Y)\bigr)$.

The score test statistic takes the form

$$\frac{1}{\bar Y(1-\bar Y)}(\mathbf{Y} - \bar Y 1)^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}(\mathbf{Y} - \bar Y 1),$$

for $\mathbf{Y}$ the vector with ith coordinate the response Y of the i individual, and $\mathbf{X}$ the matrix with ith row the regression vector $(1, X_1, \ldots, X_k)$ of the ith individual. Under the null hypothesis this possesses approximately a chisquare distribution with k degrees of freedom provided the "design matrix" $n^{-1}\mathbf{X}^T\mathbf{X}$ tends to a nonsingular matrix and the sequence $n^{-1/2}\mathbf{X}^T\mathbf{Y}$ is asymptotically normal. □

y/score	s_1	s_2	$\ldots$	s_k	
0	N_{01}	N_{02}	$\ldots$	N_{0k}	$N_{0\cdot}$
1	N_{11}	N_{12}	$\ldots$	N_{1k}	$N_{1\cdot}$
	$N_{\cdot 1}$	$N_{\cdot 2}$	$\ldots$	$N_{\cdot k}$	n

Table 9.5. Amitage test for the $(2 \times k)$ table. The expected cell frequencies are assumed to satisfy the linear model $\mathrm{E}N_{1j}/\mathrm{E}N_{\cdot j} = \alpha + \beta s_j$ for the scores given in the first column.

9.27 Example (Armitage's trend test). The *Armitage trend test* was originally proposed to investigate a linear trend in the cell frequencies in a $(2 \times k)$-table as illustrated in Table 9.5. The table could refer to a multinomial vector of order $n = \sum_{i,j} N_{ij}$, to k binomial variables (when the column totals $N_{\cdot j}$ are fixed), or two multinomial vectors of length k (when the row totals $N_{0\cdot}$ and $N_{1\cdot}$ are fixed, as in the present case-control setting). The test investigates equality in distribution of the k column vectors in the situation that it is known that the relative frequencies in the second row satisfy the model $\mathrm{E}N_{1j}/\mathrm{E}N_{\cdot j} = \alpha + \beta s_j$, for known "column scores" $s_1, \ldots, s_k$. The test is often applied in the situation that it is only known that the frequencies are ordered in size, with the scores set equal to $1, \ldots, k$. The test statistic is based on the slope coefficient in the linear regression model with n observations (X^i, Y^i), one for each individual in the table, with X^i equal to the score of the column of the individual and Y^i equal to 0 or 1 corresponding to the row.

The test can also be understood as the score test in the logistic regression model with interecept and a one-dimensional independent variable X taking the scores as its values:

$$P(Y = 1|\, X = s_j) = \frac{1}{1 + e^{-\alpha - \beta s_j}}.$$

The score test for $H_0\colon \beta = 0$ is given in Example 9.26, where we must take the matrix $\mathbf{X}$ equal to the $(n \times 2)$-matrix with ith row the vector $(1, X^i)$, for $i = 1, \ldots, n$. The test statistic can be computed to be

$$\frac{1}{\bar{Y}(1-\bar{Y})} \frac{\bigl(\sum_{i=1}^n (Y^i - \bar{Y})X^i\bigr)^2}{\sum_{i=1}^n (X^i - \bar{X})^2} = \frac{n^2}{N_{0\cdot}N_{1\cdot}} \frac{\bigl(\sum_{j=1}^k N_{1j}(s_j - \bar{s})\bigr)^2}{\sum_{j=1}^k N_{\cdot j}(s_j - \bar{s})^2},$$

where $\bar{s} = \bar{x} = \sum_{j=1}^k (N_{\cdot j}/n) s_j$ is a weighted mean of the scores. Because the hypothesis is one-dimensional we may also take the scaled score itself as the test-statistic, rather than its square length. This is the signed root of the preceding display. □

To apply the logistic model for association testing, the genotypic information on an individual must be coded in a numerical vector X. There are several possibilities to set this up:

(i) Genotypic marker mapping. The observed marker data, typically unordered genotypes at one or more marker loci, are mapped in the regression variables in a simple, direct manner.

(ii) Haplotypic marker mapping. The regression vector X is defined as a function of the individual's two haplotypes spanning several marker loci.

(iii) Causal mapping. The regression vector is defined as a function of the individual's two haplotypes spanning the (putative) causal loci.

The third possibility is attractive from a modelling perspective, but it also creates the greatest technical difficulties, as it requires a model connecting marker loci to causal loci, the genotypes at the latter being unobserved. If combined with simple ad-hoc models, the resulting procedures may be identical to the tests resulting from possibilities (i) or (ii). If the phenotype depends on the set of alleles at each of the relevant loci, rather than on their configuration in haplotypes, then strategy (ii) seems unnecessarily complicated. This assumption of absence of “cis effects” is commonly made, and seems biologically plausible for distant loci, but not for e.g. SNPs within a single gene.

Strategies (ii)-(iii) lead to logistic regression models in which the independent variable X is not observed. The score test can be adapted to this situation by conditioning the score function on the observed data, before forming the quadratic form (9.25). In fact, the score function for the model in which the observed data is a function (O,Y) of the “full data” (X,Y) is

$$\mathrm{E}_\theta\bigl(\dot\ell_\beta(Y|\,X)|\,O,Y\bigr)=\mathrm{E}_\theta\bigl((Y-\Psi(\beta^TX))X|\,O,Y\bigr).$$

This is computable from the conditional distribution of X given O. The unknown parameters θ in this distribution (for instance haplotype frequencies) are typically estimated separately from the testing procedure. Next the maximum likelihood estimator of β under the null hypothesis is computed either by setting the sum over the data of null scores equal to zero, or by the EM-algorithm??

9.28 Example (Full null hypothesis, continued). In Example 9.26 the maximum likelihood estimator of $\Psi(\beta^TX)$ under the null hypothesis is $\bar Y$, and does not depend on X. It is still the maximum likelihood estimator if X is only partially observed, and therefore the conditioned score function becomes $(Y-\bar Y)\mathrm{E}(X|\,O)$. □

9.29 Example (Single marker locus). Unordered genotypic information on a single biallelic marker locus with alleles A_1 and A_2 can assume three different values A_1A_1, A_1A_2, A_2A_2. The usual numerical coding X for these values is the number of alleles A_2 at the locus: X assumes the values 0, 1, 2 for the three genotypes. The logistic regression model says that

$$P(Y=1|\,X)=\frac{1}{1+e^{-\beta_0-\beta_1X}}.$$

The parameter β_1 models the effect of the locus on the trait, while β_0 corresponds to the prevalence of the disease in the population. The score test for testing $H_0{:}\,\beta_1=0$ is equivalent to the Armitage trend test in the (2×3)-table with the three columns equal to the numbers of controls and cases with 0, 1 and 2 alleles A_2, coded by the score $s_j=j$ for $j=0,1,2$.

The coding of the three genotypes as $0,1,2$ corresponds to an additive model, in which each allele A increases the log odds by the constant value β_1. This is not unnatural, but it does imply a choice. For instance, dominant and recessive models would use the codings 0, 1, 1 and 0, 0, 1, respectively. Because X enters the model through a linear function, two different codings will produces different outcomes unless they are linear transformations of each other. For testing $H_0{:}\,\beta_1=0$ this does not concern the level, but may influence the power of the test.

If there is no a-priori reason to assume a particular genetic model, then it may be better to use a an extra parameter instead. One possibility would be to denote the number of alleles A_2 by X_1 and to define a second regressor X_2 to be 1 for the genotype A_2A_2 and 0 otherwise. The logistic regression model becomes

$$P(Y=1|\,X)=\frac{1}{1+e^{-\beta_0-\beta_1X_1-\beta_2X_2}}.$$

The extra parameter β_2 can be interpreted as modelling a dominance effect. However, because there are three different genotypes and three parameters, the model is equivalent to permitting a different probability $P(Y=1|\,X=x)$ for each of three genotypes x, written in the forms $\Psi(\beta_0)$, $\Psi(\beta_0+\beta_1)$ and $\Psi(\beta_0+2\beta_1+\beta_2)$, respectively. □

9.30 Example (Two biallelic loci; genotypic modelling). Unordered genotypic information on a biallelic marker locus, with alleles A_1 and A_2 on the first locus and alleles B_1 and B_2 on the second, can assume nine different values, as illustrated in Table 9.1. We could define X_1 and X_2 to be the numbers of alleles A_2 at the first and B_2 at the second locus, and consider the logistic regression model

$$P(Y=1|\,X_1,X_2)=\frac{1}{1+e^{-\beta_0-\beta_1X_1-\beta_2X_2-\beta_3X_1X_2}}.$$

The parameters β_1 and β_2 are the *main effects* of the two loci, while β_2 is an *interaction effect* or in genetic terms an *epistasis effect.*

Even more so than in the preceding example the numerical coding of the three genotypes constitutes a particular modelling choice. Not only are the main effects additive, but also the interaction effect is modelled by X_1X_2, which can assume the values 0, 1, 2 and 4. Even linear transformations of the coding values 0, 1, 2 for the variables X_1 and X_2 would change the model.

We may set the epistasis parameter a-priori equal to zero, or enlarge the model with extra parameters, for instance for modelling dominance effects. The given model has three parameters, while a completely saturated model has nine free parameters. □

9.31 Example (Two loci; haplotypic modelling). For two marker loci with k and l alleles, respectively, there are kl possible haplotypes. One possible coding is to define $X_1,\ldots,X_{kl}$ as the number of haplotypes (0, 1, or 2) of each type carried by an individual. Because $\sum_j X_j=2$, we then fit a logistic regression model without intercept.

Because typically the phase is not observed, the variables $X_1,\ldots,X_{kl}$ may not be observed. In fact, the phase can be resolved, except in the case that the genotypes at both loci are heterozygous (cf. Section 9.1.2). In the latter case we replace $X_1,\ldots,X_{kl}$ by their conditional expectations given the observed genotypes. This comes down to resolving the pair of heterozygous genotypes $\{A_i,A_j\},\{B_u,B_v\}$ into the pair of haplotypes (9.4) by the formulas (9.6).

For two biallelic loci the model may be compared to the model in Example 9.30, which is defined directly in terms of the observed genotypes. With the interaction term included the latter model has four parameters, just as the present model (with

$k = l = 2$). The parameterizations are displayed in Table 9.6. In the biallelic case there are four different haplotypes, and ten different unordered pairs of haplotypes; three different unordered genotypes per locus and nine different combinations of unordered genotypes. The models are displayed relative to the latter combinations laid out in a (3×3)-table, where the central cell corresponds to two different unordered pairs of haplotypes. □

	$\{B_1, B_1\}$	$\{B_1, B_2\}$	$\{B_2, B_2\}$
$\{A_1, A_1\}$	β_0	$\beta_0 + \beta_2$	$\beta_0 + 2\beta_2$
$\{A_1, A_2\}$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1 + 2\beta_2 + 2\beta_3$
$\{A_2, A_2\}$	$\beta_0 + 2\beta_1$	$\beta_0 + 2\beta_1 + \beta_2 + 2\beta_3$	$\beta_0 + 2\beta_1 + 2\beta_2 + 4\beta_3$
$\{A_1, A_1\}$	$2\beta_1$	$\beta_1 + \beta_2$	$2\beta_2$
$\{A_1, A_2\}$	$\beta_1 + \beta_3$	$\beta_1 + \beta_4$ or $\beta_2 + \beta_3$	$\beta_2 + \beta_4$
$\{A_2, A_2\}$	$2\beta_3$	$\beta_3 + \beta_4$	$2\beta_4$

Table 9.6. Comparison of genotypic (top) and haplotypic (bottom) mapping for two biallelic loci.

9.2.4 Whole Genome Analysis

Testing for association of many marker loci, or of many groups of marker loci, requires a correction for multiple testing. Unlike in linkage analysis, there do not appear to be reliable approximations to the joint distributions of test statistics. While for linkage tests the joint distribution can be derived from stochastic models for meiosis, in association testing the distribution the distribution depends on the distribution of haplotypes in the population under study. Much data would be needed to estimate these highly-dimensional objects. HapMap Project??

For this reason multiple testing corrections are often based on general purpose methods, such as the Bonferroni correction or randomization.

9.2.5 Population Structure and Confounding

In Lemma 9.1 it was seen that two loci may well be associated in a full population, even though they are in linkage equilibrium in two subpopulations. Spurious correlation of a marker with a causal locus will generally lead to the mistaken belief that the genome near the marker is involved in causing the affection.

Besides population structure there may be other factors, generally referred as *confounders*, that cause spurious correlations. It is important to try and correct an association study for such confounding variables.

If the confounding variables were known and low-dimensional, then the most obvious method would be to perform the analysis separately for every "population" defined by a common value of the confounders. Separate p-values for the subpopulations could be combined in a single one by standard methods. In reality confounders may assume many values and/or may not be observed directly. Then corrections for confounding usually take some form of regression or mixture modelling.

In genomewide association studies an attractive possibility is to use the SNP-values themselves to infer hidden population structure. The idea is that subpopulations will be characterized by certain patterns of genomic values, determined by different allele and haplotype frequencies.

9.2.6 Genomic Control

It seems reasonable to expect that most markers in a genome-wide study are not linked to a causal locus. In that case, and in the absence of confounding, most of the test statistics for testing the null hypothesis that a given, single marker is not associated to the disease can be considered as realizations of a variable drawn from the null distribution. A deviation of the empirical distribution of all test statistics from this null distribution may be taken as a sign of spurious association by confouding.

We may hope that the values of the test statistics attached to the markers that are linked to causal loci still stand out from the those of the spuriously associated markers. Then it would work to raise the critical value of the test to eliminate the spurious markers, or, equivalently, to reduce the value of the test statistics. *Genomic control* is the procedure to divide every statistic by the quotient of the median of all test statistics and the median of the presumed null distribution.

9.2.7 Principal Components

If a population has substructure that is visible from the marker values themselves, then the principal components of the distribution of the SNPs may reveal them. It has been suggested to substract the projection on the span of the first few principal components from the genomic values before performing a case-control test.

Specifically, let $G = (G_j^i)$ be a matrix of measurements on a set of biallelic markers, coded numerically, for instance by $0, 1$ and 2 for the three possible genotypes at a single biallelic locus. The matrix has a row for every individual in the case and control groups and a column for each SNP. We view the columns as a sample from a distribution in $\mathbb{R}^n$, for n the size of the control group, and we define $a_1, \ldots, a_l$ as the eigenvectors of the empirical covariance matrix of this sample corresponding to the l largest eigenvalues, scaled to have norm 1. The columns of G, as well as the $0-1$ vector y giving the case-control status of the individual can be projected on the orthocomplement of the linear space spanned by these eigenvectors, giving a corrected matrix $\tilde{G}$ and phenotypic vector $\tilde{y}$. Next we regress $\tilde{y}$ on each column of $\tilde{G}$ and test for significance of the regression coefficient, for instance simply based on the correlation coefficient.

9.2.8 Model Selection

10
Combined Linkage and Association Analysis

In this chapter we consider some methods for gene-finding that are often classified as association methods, but do carry an element of linkage, because they make use of related individuals. The advantage of these methods over straight association is that give automatic control of confounding. The disadvantage is that they require genotyping of more individuals.

10.1 Transmission Disequilibrium Test

Suppose we sample n random individuals from the population of affected individuals, and investigate the genotypes of these individuals and their parents at a given marker locus. If the marker locus has two possible alleles, then each of the $2n$ parents can be scored in a two-way classification by type of allele transmitted to their offspring times allele nontransmitted. More precisely, each parent has an unordered genotype $\{M_1, M_1\}$, $\{M_1, M_2\}$ or $\{M_2, M_2\}$ at the marker location, and is counted in one of the four cells in the (2×2)-table in Table 10.1. For a homozygous parent transmitted and nontransmitted alleles are identical; these parents are counted in cells $(1, 1)$ or $(2, 2)$. A heterozygous parent with genotype $\{M_1, M_2\}$ is counted in cell $(2, 1)$ if he or she transmits allele M_2 and is counted in cell $(1, 2)$ if he or she transmits allele M_1. This gives a total count of $A + B + C + D = 2n$ parents in Table 10.1.

Warning. If father, mother and child are all heterozygous $\{M_1, M_2\}$, then the appropriate cell cannot be resolved for the parents individually (unless parental origins can be established). However, such a trio does contribute a count of one in both cell $(1, 2)$ and cell $(2, 1)$. Hence the pair of a father and mother can be unequivocally assigned. Table 10.2 gives examples of genotypes of trios of parents and child together with their scoring.

The total counts in the four cells of the table clearly depend both on the

		nontransmitted	
		M_1	M_2
transmitted	M_1	A	C
	M_2	B	D

Table 10.1. Transmission Disequilibrium Test. The number C is the number of parents with marker genotype $\{M_1, M_2\}$ who segregate marker allele M_1 to their child.

father	mother	child	A	B	C	D
$\{M_1, M_1\}$	$\{M_1, M_1\}$	$\{M_1, M_1\}$	2	0	0	0
$\{M_1, M_1\}$	$\{M_1, M_2\}$	$\{M_1, M_1\}$	1	0	1	0
$\{M_1, M_1\}$	$\{M_1, M_2\}$	$\{M_1, M_2\}$	1	1	0	0
$\{M_1, M_2\}$	$\{M_1, M_2\}$	$\{M_1, M_1\}$	0	0	2	0
$\{M_1, M_2\}$	$\{M_1, M_2\}$	$\{M_1, M_2\}$	0	1	1	0

Table 10.2. Examples of scoring parents for the TDT table. A trio of the given type contributes the counts listed in the columns A, B, C, D in the corresponding cell of the TDT table.

frequencies of the alleles M_1 and M_2 in the population and the relationship between the affection and the marker locus. However, if the affection has nothing to do with the marker locus, then we would expect that heterozygous parents $\{M_1, M_2\}$ transmit M_1- and M_2-alleles with equal probabilities to their (affected) children. In other words, we expect that the number of entries in the off-diagonal cells B and C are of comparable magnitude.

The *transmission disequilibrium test (TDT)* formalizes this idea by rejecting the null hypothesis of no linkage if B is large relative to $B + C$. The test may be remembered as a test for the null hypothesis that given the total number $B + C$ of heterozygous parents the number of heterozygous parents who transmit allele M_2 is binomially distributed with parameters $B + C$ and $\frac{1}{2}$. Under this binomial assumption the conditional mean and variance of B given $B + C$ are $(B + C)\frac{1}{2}$ and $(B + C)\frac{1}{2}(1 - \frac{1}{2})$, respectively. The TDT rejects the null hypothesis if

$$\frac{B - (B + C)/2}{\sqrt{(B + C)\frac{1}{4}}} = \frac{B - C}{\sqrt{B + C}}$$

is small or large relative to the standard normal distribution. Equivalently, if the square of this expression exceeds the appropriate upper quantile of the chisquare distribution with one degree of freedom.

The binomial assumption is actually not defendable, as a detailed look at the distribution of the (2×2)-table will reveal. However, we shall show that the asymptotic distribution as $n \to \infty$ of the test statistic is standard normal, identical to what it would be under the binomial assumption. We start by computing the probabilities that a given parent of an affected child gives a contribution to the cells in Table 10.1. Assume first that the affection is caused by a single, biallelic locus, and let D be the linkage disequilibrium between disease and marker loci, as defined in (9.9) and Lemma 9.10.

10.1 Lemma. *Assume that the affection is caused by a single, biallelic locus and the marker locus is biallelic with alleles M_1 and M_2. If the population is in Hardy-Weinberg equilibrium at the haplotype level, then the probability that a parent of an affected child has unordered genotype $\{M_i, M_j\}$ and transmits marker allele M_i to the child is given in cell (i, j) of Table 10.3 ($i, j \in \{1, 2\}$).*

		nontransmitted	
		M_1	M_2
transmitted	M_1	$q_1^2 + Bq_1D$	$q_1q_2 + B(q_2 - \theta)D$
	M_2	$q_1q_2 + B(\theta - q_1)D$	$q_2^2 - Bq_2D$

Table 10.3. Probability that an arbitrary parent contributes a count to the TDT table. The parameters q_1 and p_1 are the population frequencies of the marker allele M_1 and the disease allele D_1, θ is the recombination fraction between marker and disease locus, and $B = P(A)^{-1}[p_1(f_{1,1} - f_{2,1}) + p_2(f_{2,1} - f_{2,2})]$ for $f_{r,s}$ the penetrances of disease genotype (D_r, D_s) and $P(A)$ the prevalence of the disease.

In order to find a gene that causes the affection we would like to test the null hypothesis $H_0{:}\,\theta = \frac{1}{2}$ that the marker locus is unlinked to the disease locus. Under this hypothesis the off-diagonal probabilities in Table 10.3 reduce to the same value

$$q_1q_2 + B(\tfrac{1}{2} - q_1)D = q_1q_2 + B(q_2 - \tfrac{1}{2})D.$$

Thus the hypothesis of no linkage implies that the (2×2)-table is symmetric, and it makes sense to perform a test for equality of the off-diagonal probabilities. This is exactly what the TDT aims for. Furthermore, the TDT appears justified to use only the variables B and C from Table 10.1, as the expected values of the the diagonal elements do not depend on θ.

The two off-diagonal probabilities are also equal if $D = 0$, irrespective of the value of the recombination fraction θ. It follows that the TDT can have power as a test for $H_0{:}\,\theta = \frac{1}{2}$ only if $D \neq 0$. This is often expressed by saying "that the TDT is a test of linkage (only) if the disease and marker loci are associated", and is a mathematical expression of the observation in Chapter 9 that a case-control approach can be successful only if there is linkage disequilibrium between the marker and causal locus. It is clear also that if association is present, but very small ($D \approx 0$), then the TDT will have some power to detect linkage, but too little to give conclusive results with not too large data-sets.

The fact that the TDT is a correct test for $H_0{:}\,\theta = \frac{1}{2}$ for any $D > 0$ indicates that it can correctly handle "spurious association", such as caused by population admixture.[†]

Table 10.1 gives a two-way classification of $2n$ individuals, and it is tempting to view (A, B, C, D) as the realization of a multinomial vector with parameters $2n$ and the four probabilities in Table 10.3. This is wrong, because the $2n$ individuals are not independently classified: even though the n families can be assumed to form

[†] Need to remove the Hardy-Weinberg type assumptions of the theorem to make this a valid argument??

independent units, the status of the father and mother of an affected child within a family are not independent in general. This is due to the fact that the trios of father, mother and child are selected based on the information that the child is affected. If the marker is associated to the disease, then the information that the child is affected together with information about the marker allele transmitted by one parent provides information about the marker allele transmitted by the other parent. For instance, if the first parent did not transmit the disease allele, then the second parent probably did, and this is informative about the marker allele transmitted. Thus we cannot use a multinomial model for Table 10.1. It is also not true in general that under the null hypothesis of no linkage the variable B is conditionally binomially distributed given $B+C$. (In the absence of association the contributions of the parents within families to the table are indeed independent, and the conditional distribution of B given $B+C$ is binomial. However, this is of no use, because given zero association, there is no power to detect linkage, and we would not want to carry out the TDT in the first place.) However, it is shown in the following lemma that the normal approximation is nevertheless correct.

A correct approach is to take the n family trios as independent sampling units, instead of the $2n$ parents. This is possible by replacing the (2×2)-table by a (4×4)-table, whose 16 cells register the transmission patterns of the n father-mother pairs: on each axis we place the four patterns M_1M_1, M_1M_2, M_2M_1 and M_2M_2, indicating pairs of a parental and maternal allele, transmitted (one axis) or nontransmitted (other axis). The statistical model can then be summarized by saying that the (4×4)-table is multinomial with parameters n and 16 probabilities. These probabilities are expressed in the parameters of the underlying disease model in Lemma 10.3. The probabilities in Table 10.3 can be derived from them.

The (4×4)-table gives complete information on the distribution of the observations underlying the TDT test. We may still decide to base the test on the difference $B-C$ of the off-diagonal elements in Table 10.1, scaled to an appropriate standard distribution. The following lemma shows that the scaling of the TDT (badly motivated previously by the assumption of a binomial distribution) gives the correct significance level, at least for large n.

10.2 Lemma. *Under the assumptions of Lemma 10.1 the sequence of variables $(B-C)/\sqrt{B+C}$ tends to a standard normal distribution under the null hypothesis $H_0\colon \theta = \frac{1}{2}$, as $n\to\infty$.*

Proof. Distributions, expectations or variances in this proof will silently be understood to be conditional on the null hypothesis and on the event A that the sib is affected.

Let N^P and N^M be the (2×2)-tables contributed to the TDT-table Table 10.1 by the father and mother of a typical family trio, so that $N = N^P + N^M$ is the total contribution of the family trio, and the variable $B-C$ is the sum of the n variables $N_{12}-N_{21}$ contributed by the n families. The key to the lemma is that the variables $N_{12}^P - N_{21}^P$ and $N_{12}^M - N_{21}^M$ are uncorrelated under the null hypothesis, even though dependent in general.

The variable N^P_{ij} is equal to 1 or 0 according to the occurrence or nonoccurrence of the event $M^P_{i,j}\cap T^P_i$ as described in Table 10.4; the same is true for N^M_{ij} relative to the event $M^M_{ij}\cap T^M_j$. The probabilities of the two events $M^P_{12}\cap T^P_1$ and $M^P_{12}\cap T^P_2$ are the off-diagonal elements in the TDT table Table 10.3, and are equal under the null hypothesis $H_0{:}\,\theta=\frac12$. The same is true for the maternal contributions. It follows that $\mathrm{E}_0(N^P_{12}-N^P_{21}|\,A)=\mathrm{E}_0(N^M_{12}-N^M_{21}|\,A)=0$.

The product $(N^P_{12}-N^P_{21})(N^M_{12}-N^M_{21})$ is equal to 1 if both terms are 1 or both terms are -1, and it is -1 otherwise. It follows that the covariance of the variables $N^P_{12}-N^P_{21}$ and $N^M_{12}-N^M_{21}$ is given by

$$\begin{aligned}&\mathrm{E}_0\bigl((N^P_{12}-N^P_{21})(N^M_{12}-N^M_{21})|\,A\bigr)\\&\qquad=P_0\bigl(M^P_{12}\cap T^P_1\cap M^M_{12}\cap T^M_1|\,A\bigr)+P_0\bigl(M^P_{12}\cap T^P_2\cap M^M_{12}\cap T^M_2|\,A\bigr)\\&\qquad\quad-P_0\bigl(M^P_{12}\cap T^P_1\cap M^M_{12}\cap T^M_2|\,A\bigr)-P_0\bigl(M^P_{12}\cap T^P_2\cap M^M_{12}\cap T^M_1|\,A\bigr).\end{aligned}$$

The probabilities on the right are given in Lemma 10.3. With the notation $p_{ij,r}=(1-\theta)h_{ri}q_j+\theta h_{rj}q_i$, the expression on the right can be written in the form

$$\frac{1}{P(A)}\sum_r\sum_s f_{r,s}\bigl[p_{12,r}p_{12,s}+p_{21,r}p_{21,s}-p_{12,r}p_{21,s}-p_{21,r}p_{12,s}\bigr].$$

Under the null hypothesis $H_0{:}\,\theta=\frac12$ we have that $p_{ij,r}=p_{ji,r}$, and hence the expression in the display vanishes. This concludes the proof that the contributions of fathers and mothers to the TDT-table are uncorrelated.

The variable N^P has a (2×2)-multinomial distribution with parameters 1 and (2×2) probability matrix given in Table 10.1. Under the null hypothesis the off-diagonal elements of the probability matrix are equal, say c. Then

$$\begin{aligned}\mathrm{E}_0(N^P_{12}|\,A)&=\mathrm{E}_0(N^P_{21}|\,A)=c,\\\mathrm{var}_0(N^P_{12}-N^P_{21}|\,A)&=\mathrm{var}_0(N^P_{12}|\,A)+\mathrm{var}_0(N^P_{21}|\,A)-2\,\mathrm{cov}_0(N^P_{12},N^P_{21}|\,A)\\&=2c(1-c)-2(0-c^2)=2c.\end{aligned}$$

The maternal versions of these variables satisfy the same equalities.

The variable $B-C$ is the sum over families of the variables $N^P_{12}-N^P_{21}$ and $N^M_{12}-N^M_{21}$. The mean of $B-C$ is zero. Because contributions of families are independent and contributions of fathers and mothers uncorrelated, the variance of $B-C$ is equal to $2n2c$. The independence across families allows to apply the central limit theorem and yields that the sequence $(B-C)/\sqrt{2n2c}$ tends in distribution to a normal distribution as $n\to\infty$. The independence across families and the law of large numbers gives that $(B+C)/n$ tends in probability to $4c$. Together these assertions imply the lemma, in view of Slutsky's lemma. ∎

We can also consider the TDT as a test for the null hypothesis $H_0{:}\,D=0$ of no association. Because for $\theta=\frac12$, the off-diagonal probabilities in Table 10.3 are the same no matter what the value of D, the TDT will have power only if the marker and disease loci are linked. Under the null hypothesis the vector (A,B,C,D) is multinomially distributed with probability vector $(q_1^2,q_1q_2,q_1q_2,q_2^2)$.

event	*interpretation*
A	child is affected
M_{ij}^P	father has marker genotype $\{M_i, M_j\}$
M_{ij}^M	mother has marker genotype $\{M_i, M_j\}$
T_i^P	father transmits marker allele M_i
T_i^M	mother transmits marker allele M_i
D_r^P	father transmits disease allele D_r
D_r^M	mother transmits disease allele D_r

Table 10.4. Events used in the proof of Lemma 10.3.

* 10.1.1 Multiple Alleles

The TDT can be extended to marker and disease loci with more than two possible alleles. Consider a marker locus with possible alleles $M_1, \ldots, M_k$, and a monogenetic disease that is caused by a gene with alleles $D_1, \ldots, D_l$. Then any parent can be classified according to a $(k \times k)$-table, contributing a count to cell (i, j) if the parent possesses unordered genotype $\{M_i, M_j\}$, transmits allele M_i and does not transmit allele M_j to the child. A combination of the transmission data of a father and a mother can be classified in a $(k^2 \times k^2)$-table. The probabilities of the latter table are given in the following lemma. Let h_{ij} and q_j be the frequencies of haplotype D_iM_j and marker allele M_j in the general population, respectively, and let $f_{r,s}$ be the probability of affection given the (ordered) genotype (D_r, D_s).

10.3 Lemma. *Assume that the disease is monogenetic. If the population is in Hardy-Weinberg equilibrium at the haplotype level, then the conditional probability given the event A that a father has unordered marker genotype $\{M_i, M_j\}$ and transmits allele M_i and the mother has unordered marker genotype $\{M_u, M_v\}$ and transmits allele M_u is equal to*

$$\frac{1}{P(A)} \sum_r \sum_s f_{r,s} \big((1-\theta)h_{ri}q_j + \theta h_{rj}q_i\big)\big((1-\theta)h_{su}q_v + \theta h_{sv}q_u\big). \tag{10.4}$$

Here $P(A)$ is the prevalence of the disease in the population and $f_{r,s}$ is the penetrance of the disease genotype (D_r, D_s).

Proofs. With the notation for events given in Table 10.4 the event of interest is $M_{ij}^P \cap T_i^P \cap M_{uv}^M \cap T_u^M$. The conditional probability of this event can be written

$$\begin{aligned} &P(M_{ij}^P \cap T_i^P \cap M_{uv}^M \cap T_u^M \,|\, A) \\ &\quad = \sum_r \sum_s P(M_{ij}^P \cap T_i^P \cap D_r^P \cap M_{uv}^M \cap T_u^M \cap D_s^M \,|\, A) \\ &\quad = \frac{1}{P(A)} \sum_r \sum_s f_{r,s} P(M_{ij}^P \cap T_i^P \cap D_r^P) P(M_{uv}^M \cap T_u^M \cap D_s^M), \end{aligned}$$

by Bayes' formula. The probabilities for the paternally and maternally determined events on the far right are identical. For simplicity of notation we drop the superscript P or M. The events can be decomposed on the occurrence of a crossover between the disease and marker locus. In the absence of a crossover the event $M_{ij}\cap T_i\cap D_r$ occurs if the parent has haplotype D_rM_i on one chromosome, marker allele M_j on the other chromosome and passes on the marker allele M_i, which has conditional probability $2h_{ri}q_j\frac{1}{2}$. Given a crossover the event $M_{ij}\cap T_i\cap D_r$ occurs if the parent has haplotype D_rM_j on one chromosome, marker allele M_i on the other chromosome and passes on marker allele M_i, which has probability $2h_{rj}q_i\frac{1}{2}$. Thus $P(M_{ij}\cap T_i\cap D_r)=(1-\theta)h_{ri}q_j+\theta h_{rj}q_i$, and the right side of the preceding display reduces to formula (10.4). This concludes the proof of Lemma 10.3.

Under Hardy-Weinberg equilibrium the ordered disease genotype (D_r,D_s) has probability p_rp_s and hence $P(A)=\sum_r\sum_s f_{r,s}p_rp_s$.

By marginalizing the haplotype frequencies we have $\sum_j h_{ij}=p_i$. Therefore, marginalization of formula (10.4) over u and v readily yields that

$$P(M_{ij}^P\cap T_i^P\,|\,A)=\frac{1}{P(A)}\sum_r\sum_s f_{r,s}\big((1-\theta)h_{ri}q_j+\theta h_{rj}q_i\big)p_s. \tag{10.5}$$

The left side gives the probabilities described in Table 10.3, if specialized to biallelic loci and $i,j\in\{1,2\}$. To prove Lemma 10.1 we need to show that the right sides can be written in the form as given in the table.

In the absence of association between marker and disease locus the haplotype frequencies satisfy $h_{ij}=p_iq_j$ and hence $(1-\theta)h_{ri}q_j+\theta h_{rj}q_i=q_iq_jp_r$. Given this factorization for all possible pairs of alleles, the preceding display (10.5) can be seen to reduce to

$$\frac{1}{P(A)}\sum_r\sum_s f_{r,s}q_iq_jp_r\,p_s=q_iq_j.$$

We need to express the deviation from this equilibrium value in the parameters D and θ. By Lemma 9.10,

$$\begin{aligned}h_{11}&=p_1q_1+D,\\ h_{12}&=p_1q_2-D,\\ h_{21}&=p_2q_1-D,\\ h_{22}&=p_2q_2+D.\end{aligned}$$

Inserting these four representations in (10.5), we can split the resulting expression in the equilibrium value as in the preceding paragraph and an expression that is a multiple of D. Straightforward algebra shows that the latter expression reduces to the values Bq_1D, $B(q_2-\theta)D$, $B(\theta-q_1)D$ and $-Bq_2D$, for the four elements in Table 10.3, respectively. ■

Under the assumption that the n families are sampled independently, the data can be summarized as a $(k^2\times k^2)$-table with a multinomial distribution with parameters n and the k^4 probabilities in the preceding lemma. An extended TDT should test whether the probabilities of the cells in the table are symmetric across the diagonal. This can be achieved through a variety of test statistics.

As in the case of biallelic markers it is customary to pool fathers and mothers, and collaps the data into a $(k \times k)$-table, where cell (i, j) counts how many parents with unordered genotype $\{M_i, M_j\}$ transmit allele M_i and do not transmit allele M_j to their child. A natural extension of the TDT to multiallelic markers is then

$$\sum\sum_{i<j} \frac{(N_{ij} - N_{ji})^2}{N_{ij} + N_{ji}}.$$

By the same method of proof as in Lemma 10.2 the variables $N_{ij} - N_{ji}$ can be shown to be uncorrelated for different pairs (i, j), and the test statistic can be shown to be asymptotically chisquare distributed with $\binom{k}{2}$ degrees of freedom under the null hypothesis of no linkage. Unfortunately, the statistic turns out to have poor overall power, possibly because it compares all cells individually. Symmetry of the $(k \times k)$-table implies symmetry of its marginal values $N_{i\cdot} = \sum_j N_{ij}$ and $N_{\cdot i} = \sum_j N_{ji}$ and hence an alternative is a quadratic form such as

$$\sum_i \sum_j (N_{i\cdot} - N_{\cdot i}) \alpha_{ij} (N_{j\cdot} N_{\cdot j}).$$

For (α_{ij}) (an estimate of) the covariance matrix of the vector $(N_{1\cdot} - N_{\cdot 1}, \ldots, N_{k\cdot} - N_{\cdot k})$ this statistic ought to have approximately a chisquared distribution with $k - 1$ degrees of freedom under the null hypothesis. (Thus $\alpha_{ij} = -(N_{ij} + N_{ji})$ for $i \neq j$ and $N_{i\cdot} + N_{\cdot i} - 2N_{ii}$ otherwise.)

A third possible test statistic can be derived from modelling the probabilities in the table by significantly fewer parameters than the number of cells.

???? In the preceding it is assumed that the affected children belong to different families, so that the n trios contain n different pairs of parents. If we sample two affected individuals from the population and they turn out to have the same parents, then we can still form two parents-child trios, but the parents will occur twice. The contributions to the TDT-table of two of such trios are not independent. However, it appears that the contributions to the TDT statistic are uncorrelated and hence the variance of the TDT statistic is as if the trios were independent. ????

10.6 EXERCISE. Show that the TDT statistic arises as the chisquare statistic (under the (wrong) assumption that the vector (A, B, C, D) is multinomially distributed with parameter $2n$) for testing the null hypothesis that the off-diagonal probabilities are equal.

10.2 Sibship Transmission Disequilibrium Test

The TDT is based on observing a child and its parents, or children and their parents. The *sibship transmission disequilibrium test (S-TDT)* is based on observing marker data on the children only. We select a sample of sibships, each consisting of affected and unaffected children. For a biallelic marker, we measure the total number of alleles M_1 carried by the affected sibships and compare this to the number of alleles carried by the unaffected sibships. If the affected sibs carry "more" alleles M_1, then we conclude that the marker locus is linked to the affection.

More precisely, suppose that the ith sibship consists of n_A^i affected and n_U^i unaffected children, and N_A^i of the $2n_A^i$ marker alleles of the affected children are allele M_1, and N_U^i of the $2n_U^i$ marker alleles of the unaffected children are M_1. Let $n_A = \sum_i n_A^i$ and $n_U = \sum_i n_U^i$ be the total number of affected and unaffected children in the sibships, and let $N_A = \sum_i N_A^i$ and $N_U = \sum_i N_U^i$ be the total numbers of alleles M_1 carried by these two groups. If N_A/n_A is significantly different from N_U/n_U, then this is an indication that the marker locus is associated with the disease locus, and possibly linked.

To make "significantly different" precise, it would be nice if we could assume that N_A and N_U are independent binomial variables with parameters (n_A, p_A) and (n_U, p_U). We would then test the null hypothesis that $H_0\colon p_A = p_U$. However, the selection of sibships rather than inviduals renders this hypothesis untrue. We can carry out a permutation test instead.

Because under the null hypothesis marker and affection are unrelated, each redistribution of affection status within a sibship is equally likely. Given the numbers n_A^i and n_U^i and the distribution of the marker alleles M_1 over the $n_A^i + n_U^i$ sibs in a sibships, we might reassign the labels "affected" or "unaffected" in a random manner by choosing n_A^i arbitrary sibs to be affected and the remaining sibs to be unaffected. We perform this independently across all sibships. Under this reassignment the variables N_A/n_A and N_U/n_U assume new values, which we may assume as random as long as we say that we randomly reassign the affection lables. Thus we create probability distributions for these variables, their "permutation distributions". We determine a critical value from these permutation distributions.

In practice, we generate as many random reassigments of the affection labels as computationally feasible. For each reassigment we calculate the corresponding value of N_A/n_A. If the value of N_A/n_A on the real data is among the 5 % most extreme values, then we reject the null hypothesis.

Of course, there are other possible test statistics. For instance, the "SDT" uses

$$\#\Big(i\colon \frac{N_A^i}{n_A^i} > \frac{N_U^i}{n_U^i}\Big).$$

*11
Coalescents

Any present-day allele has been segregated by a parent in a previous generation. A single-locus allele (which is not recombined) can be traced back to a single allele in the previous generation of alleles, and by iteration to an ancestor in all previous generations of alleles. Two different alleles may have the same parent allele, and if we trace back in time far enough, then we shall find that any given set of alleles descend from a single parent allele somewhere in the past. The first such parent is called a *most recent common ancestor* or MRCA. The *coalescent* is a stochastic process for the tree structure describing the inheritance process. After adding mutation and recombination it can be used to map the origins of disease genes, or estimate the age of an allele: the time that it first arose.

In this chapter we consider Kingman's coalescent, which is model for *neutral alleles*, in the sense that it does not take into account evolutionary selection.

11.1 Wright-Fisher Model

Consider a population of N individuals, labelled arbitrarily with the symbols $1, 2, \ldots, N$. Suppose that individual i has M_i descendants, where the vector $(M_1, \ldots, M_N)$ is multinomially distributed with parameters N and $(1/N, \ldots, 1/N)$. Thus the original population is replaced by a new population of N children. We label the new population with the symbols $a_1, a_2, \ldots, a_N$ in a random order.

Note that in this simple model the children are born without mating, a child has only one parent, and a fortiori a process of recombination of chromosomes is absent.

From symmetry considerations it is evident that a given child a_i has a given parent j with probability $1/N$. (Alternatively, see the first part of the proof of the following lemma.) A bit reversing reality, it is said that *a child chooses its parent at random from the previous generation*. The following lemma shows that the children choose their parents independently.

11.1 Lemma. *The probability that children $a_{i_1},\ldots,a_{i_k}$ choose parents $j_1,\ldots,j_k$ is equal to $(1/N)^k$, for any set of different $\{i_1,\ldots,i_k\}\subset\{1,\ldots,N\}$ and any choice $j_1,\ldots,j_k\in\{1,\ldots,N\}$.*

Proof. The multinomial distribution with parameters N and $(p_1,\ldots,p_k)$ is the distribution of the numbers of balls in k given boxes if N balls are placed independently and at random in the k boxes. Therefore, if the N children choose their parents independently and at random from N given parents, then the number of children of the N parents is a multinomial vector with parameters N and $(1/N,\ldots,1/N)$. This proves the lemma without computations.

We can also prove the lemma using the numerical definition of the multinomial distribution as the starting point. To illustrate the idea of the computation first consider the case $k=1$. Let E be the event that child a_i has parent j. If $M_j=0$, then the (conditional) probability of E is zero, since parent j has no children in that case. Represent the N children born to the N parents by the labels of their parents. If $M_j=m$, then parent j has m children and hence these N labels include m times the symbol j. The event E occurs if the random permutation of the N symbols referring to the children has a symbol j in the ath position. Thus

$$\begin{aligned}P(E)&=\sum_{m=1}^N P(M_j=m)P(E|\,M_j=m)\\&=\sum_{m=1}^N P(M_j=m)\frac{m}{N}=\frac{1}{N}\mathrm{E}M_j=\frac{1}{N}.\end{aligned}$$

The last equality follows because M_j is binomially distributed with parameters N and $1/N$.

To prove the lemma in the general case, suppose that the parents $j_1,\ldots,j_k$ consists of n_i times parent i (with possibly $n_i=0$) for $i=1,\ldots,N$, so that the set $j_1,\ldots,j_k$ can be ordered as

$$\overbrace{1,\ldots,1}^{n_1\text{ times}},\overbrace{2,\ldots,2}^{n_2\text{ times}},\ldots\ldots,\overbrace{N,\ldots,N}^{n_N\text{ times}}.$$

Without loss of generality we can order the children so that the event E of interest becomes that the first n_1 children have parent 1, the second n_2 children have parent 2, etc. If $M_i<n_i$ for some $i=1,\ldots,N$, then the event E has probability zero. Represent the N children born to the N parents by the labels of their parents. If $M_i=m_i$ with $m_i\geq n_i$ for every i, then the event E occurs if the random ordering of m_1 symbols 1, m_2 symbols 2, etc. has the symbol 1 in the first n_1 places, the symbol 2 in the second n_2 places, etc. Thus, with the multiple sums restricted to

indices with $\sum_j m_j = N$,

$$\begin{aligned}
P(E) &= \sum_{m_1=n_1}^{N} \cdots \sum_{m_N=n_N}^{N} P(M=m)P(E|\,M_1=m_1,\ldots,M_N=m_N) \\
&= \sum_{m_1=n_1}^{N} \cdots \sum_{m_N=n_N}^{N} \frac{N!}{m_1!\cdots m_N!}\Big(\frac{1}{N}\Big)^N \\
&\qquad\qquad \times \frac{\prod_{j=1}^N m_j(m_j-1)\cdots(m_j-n_j+1)\times(N-\sum_j n_j)!}{N!} \\
&= \sum_{m_1=n_1}^{N} \cdots \sum_{m_N=n_N}^{N} \frac{(N-\sum_j n_j)!}{(m_1-n_1)!\cdots(m_N-n_N)!}\Big(\frac{1}{N}\Big)^N \\
&= \Big(\frac{1}{N}\Big)^{\sum_j n_j}.
\end{aligned}$$

This proves the theorem as $\sum_j n_j$ is the number of children involved. ∎

Because the population size is assumed to be constant, each parent is on the average replaced by one child. The probability that a given parent has no children is equal to $(1-1/N)^N$, and hence for large N on the average the lineages of Ne^{-1} of the parents die out in a single generation.

We shall study this in more detail by repeating the reproduction process a number of times, giving generations

$$\begin{gathered}
1,2,\ldots,N, \\
a_1^{(1)},a_2^{(1)},\ldots,a_N^{(1)}, \\
a_1^{(2)},a_2^{(2)},\ldots,a_N^{(2)}, \\
\vdots
\end{gathered}$$

Each individual in the kth generation is the child of an individual in the $(k-1)$th generation, which is the child of an individual in the $(k-2)$th generation, and so on. Starting with an individual $a_i^{(k)}$ in the kth generation we can thus form a chain of child-parent relationships linking $a_i^{(k)}$ to one of the parents in $\{1,2\ldots,N\}$ at time 0. Graphically, these chains can be pictured as the sample paths of N random walks (one for each individual $a_i^{(k)}$), as in Figure 11.1. The state space of the random walks (the vertical lines in the figure) is identified with the set of labels $\{1,2\ldots,N\}$, by placing individual $a_j^{(l)}$ at location j on the vertical line at time l. (Thus the state space has no spatial interpretation, unlike with an ordinary random walk.)

Two individuals may have the same parent. For instance, this is the case for individuals $a_2^{(k)}$ and $a_5^{(k)}$ in Figure 11.1, and also for individuals $a_3^{(k-2)}$ and $a_5^{(k-2)}$. The random walks containing these individuals then *coalesce* at the parent, and remain together from then on. Since the individuals choose their parents independently and at random the probability that two individuals choose the same parent,

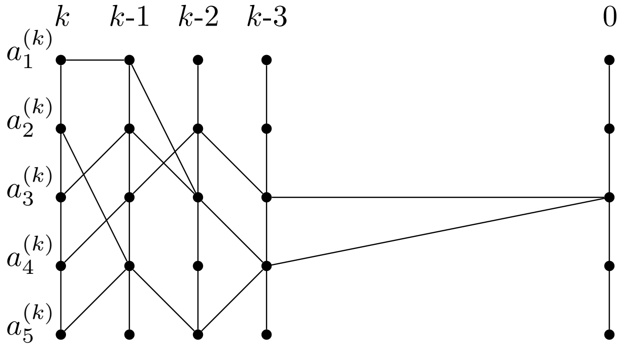

Figure 11.1. Coalescent paths for $N = 5$.

so that their random walks coalesce in the next step, is $1/N$. The probability that the random walks of two individuals remain separated for at least l generations is $(1 - 1/N)^l$.

In Figure 11.1 it is suggested that the random walks of all individuals in the kth generation have coalesced at generation 0. In other words, all individuals in the kth generation descend from a single ancestor in the 0th generation. This is unlikely if k is small relative to N, but has probability tending to 1 if $k \to \infty$ and N is fixed. The latter follows because the probability of no coalescence of two particular walks is $(1 - 1/N)^k$, so that the probability of noncoalescence of some pair of walks is certainly not bigger than

$$\binom{N}{2}\Bigl(1-\frac{1}{N}\Bigr)^k.$$

This tends to zero as $k \to \infty$ for fixed N.

Because the transition from a generation to a previous generation is completely described by the rule that childeren choose their parents independently at random and the transitions are independent across generations, the process depicted in Figure 11.1 has a natural extension to the times $-1, -2, \ldots,$. The preceding paragraph shows that eventually, if we go far enough to the right, the random walks will coalesce. The time that this happens is of course a random variable, and this takes arbitrarily small (negative) values with positive probability. The individual in which the paths coalesce is called the *most recent common ancestor (MRCA)* of the populaton.

If we thus extend time to the right, the indexing of the generations by the numbers $k, k-1, \ldots, 1, 0, -2, \ldots$ is awkward. Hence we replace it by $0, 1, 2, \ldots,$ as in Figure 11.2. Time then runs from 0 to ∞, in the reverse direction relative to natural time. We label the starting points of the random walks at time 0 by $1, 2, \ldots, N$ instead of $a_1^{(k)}, a_2^{(k)}, \ldots, a_N^{(k)}$.

At a given (reverse) time k some of the N random walks started at time 0 will

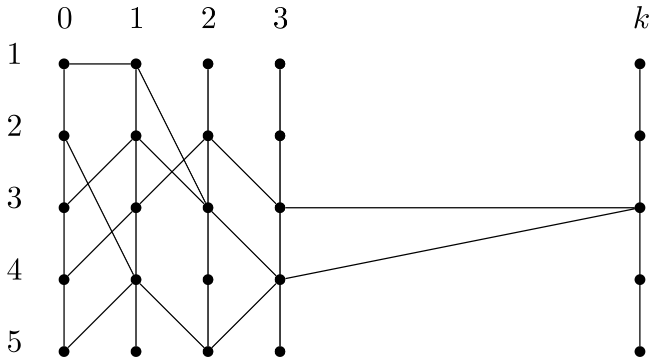

Figure 11.2. Coalescent paths.

have coalesced, while others are still separated. This induces a partition

$$\{1,2,\ldots,N\} = \cup_{i=1}^{N_k} X_k^N(i),$$

where every of the partitioning sets $X_k^N(i)$ contains the starting points of random walks that have coalesced before or at time k. (More formally, we define two random walks to be equivalent if they have coalesced before or at time k; this partitions the random walks in a set of equivalence classes; the partitioning sets $X_k(i)$ are the starting labels of the random walks in these classes.) At time 0 no walks have coalesced and hence the partition at time zero is $\cup_{i=1}^N \{i\}$. Because coalesced random walks remain together, the sequence of partitions

$$\cup_{i=1}^N \{i\}, \quad \cup_{i=1}^{N_1} X_1^N(i), \quad \cup_{i=1}^{N_2} X_2^N(i), \quad \ldots\ldots$$

are successive coarsenings: each set $X_k^N(i)$ is the union of one or more sets $X_{k-1}^N(j)$ of the previous partition. The sequence of partitions forms a stochastic process, which we denote by

$$X_0^N, X_1^N, X_2^N, \ldots.$$

The state space of this process is the collection of all partitions of the set $\{1,\ldots,N\}$. The process is Markovian with stationary transition function, as at each time the further coarsening is independent of the way the current partition was reached and is determined by the rule that childeren choose parents independently at random. The number of partitions K_N ("Bellman numbers") of $\{1,\ldots,N\}$ increases rapidly with N. The transition matrix P_N of the process is a $(K_N \times K_N)$-matrix, which is huge for large N. However, because each transition is a coarsening, most entries of the matrix P_N are zero. The partition of $\{1,\ldots,N\}$ into a single set is an absorbing state and can be reached from every other state. It follows that the Markov chain will reach this state eventually with probability 1.

	1|2|3	12|3	1|23	13|2	123
1|2|3	2/9	2/9	2/9	2/9	1/9
12|3	0	2/3	0	0	1/3
1|23	0	0	2/3	0	1/3
13|2	0	0	0	2/3	1/3
123	0	0	0	0	1

Table 11.1. Transition matrix P_N of the process $X_0^N, X_1^N, \ldots$ for $N = 3$.

11.2 EXERCISE. Verify Table 11.1.

Besides that many elements of P_N are exactly zero, for large N most nonzero entries are almost zero. The following calculations show that the most likely "transition" is to stay in the same state ("no coalescence"), and the second likely transition is the coalescence of exactly two random walks.

To compute the probability of a "constant transition" from a partition x into itself denote by $\#x$ the number of lineages (i.e. number of sets in the partition). Then at time $k+1$ the same set of $\#x$ lineages exists if all $\#x$ individuals who are the kth generation representatives of the $\#x$ lineages choose different parents. Thus

$$\begin{aligned} P(X_{k+1}^N = x \,|\, X_k^N = x) &= \frac{N(N-1)\cdots(N-\#x+1)}{N^{\#x}} \\ &= 1 - \binom{\#x}{2}\frac{1}{N} + O\Bigl(\frac{1}{N^2}\Bigr), \end{aligned} \tag{11.3}$$

as $N \to \infty$, where $\#x$ is the number of partitioning sets in x.

For given partitions x and y of $\{1, \ldots, N\}$, write $x \Rightarrow y$ if y can be obtained from x by uniting two subsets of x, leaving the other partitioning sets untouched. If $X_k^N = x$, then $X_{k+1}^N = y$ for some y with $x \Rightarrow y$ if the two kth generation representatives of the lineages that are combined in the transition $x \Rightarrow y$ choose the same parent, and the other ancestors choose different parents. The probability of this event is

$$P(X_{k+1}^N = y \,|\, X_k^N = x) = \frac{N(N-1)\cdots(N-\#x+2)}{N^{\#x}} = \frac{1}{N} + O\Bigl(\frac{1}{N^2}\Bigr), \tag{11.4}$$

where $\#x$ is the number of partitioning sets in x.

For a given partition x there are $\binom{\#x}{2}$ partitions y with $x \Rightarrow y$. From combining the preceding two displays it therefore follows readily that the probability of a transition from x to some y with $y \neq x$ and not $x \Rightarrow y$ is of the lower order $O(1/N^2)$. Staying put has probability $1 - O(1/N)$, while transitions of the form $x \Rightarrow y$ account for most of the remaining $O(1/N)$-probability. In other words, the diagonal elements of the transition matrix P_N are $1 - O(1/N)$, there are $\binom{\#x}{2}$ elements of the order $O(1/N)$ in the xth row with the same leading (nonzero) term of order $1/N$, and the remaining elements are zero or $O(1/N^2)$.

To study the limiting case as $N \to \infty$, it is inconvenient that the dimension of the state space of the process X^N becomes larger and larger as $N \to \infty$. To avoid this we shall consider only the partitions induced on a fixed set of individuals, for instance those numbered $1, \ldots, n$ for a given n. We then obtain a Markov chain $X^{n,N}$ with state space the set of partitions of the set $\{1, 2, \ldots, n\}$. The full population remains a set of N individuals, in the sense that each generation of children chooses their parents independently at random from a population of N parents. The difference is that we only follow the random walks that originate at time zero at one of the points $1, 2, \ldots, n$. The preceding reasoning applies in the same way to the processes $X^{n,N}$, i.e. equations (11.3) and (11.4) are valid with $X^{n,N}$ substituted for X^N for any partitions x, y of the the set $\{1, 2, \ldots, n\}$. (The transition matrix of $X^{n,N}$ can also be derived from the transition matrix of X^N, but this is a bit complicated.)

To study the limiting case as $N \to \infty$ for fixed n, we define a continuous time stochastic process $(Y_t^N : t \geq 0)$ by

$$Y_{j/N}^N = X_j^{n,N}, \qquad j = 0, 1, 2, \ldots,$$

and define Y_t^N to be constant on the intervals $[j/N, (j+1)/N)$. Thus the original process $X_0^{n,N}, X_1^{n,N}, X_2^{n,N}, \ldots$ become the skeleton of the process Y^N at the times $0, 1/N, 2/N, \ldots$. The process Y^N inherits the Markov property of the process $X^{n,N}$ and its state space: the collection of partitions of the set $\{1, 2, \ldots, n\}$. We shall show that as $N \to \infty$ the sequence of processes Y^N tends to a Markov process with generator matrix A, given by

$$A(x, y) = \begin{cases} -\binom{\#x}{2}, & \text{if } y = x, \\ 1, & \text{if } x \Rightarrow y, \\ 0, & \text{otherwise.} \end{cases} \tag{11.5}$$

11.6 Theorem. *The transition matrices $Q_t^{(N)}$ defined by $Q_t^{(N)}(x, y) = P(Y_{s+t}^N = y \mid Y_s^N = x)$ satisfy $Q_t^{(N)} \to e^{tA}$ as $N \to \infty$, for any $t > 0$.*

Proof. From the preceding it follows that the transition matrix P_N of the process $X^{n,N}$ satisfies, as $N \to \infty$

$$A_N := N(P_N - I) \to A.$$

This implies that, for any $t > 0$,

$$P_N^{\lfloor Nt \rfloor} = \Bigl(I + \frac{A_N}{N}\Bigr)^{\lfloor Nt \rfloor} = \sum_{k=0}^{\lfloor Nt \rfloor} \binom{\lfloor Nt \rfloor}{k} \Bigl(\frac{A_N}{N}\Bigr)^k \to \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} = e^{tA}.$$

The same is true with $\lfloor Nt \rfloor$ replaced by $\lfloor Nt \rfloor + 1$.

From the definition of Y^N it follows that the transitions of Y^N in the time interval $(s, s+t]$ consist of transitions of the process $X^{n,N}$ at the time points $k+1, k+2, \ldots, k+l$ for k, l integers determined by $k/N \leq s < (k+1)/N \leq (k+l)/N \leq s+t < (k+l+1)/N$. These inequalities imply that $Nt - 1 < l < Nt + 1$, whence there are $\lfloor Nt\rfloor$ or $\lfloor Nt\rfloor+1$ transitions during the interval $(s,t]$, and hence the transition matrix

$$\Bigl(P(Y^N_{t+s}=y|\,Y^N_s=x)\Bigr)$$

is given by $P_N^{\lfloor Nt\rfloor}(x,y)$ or $P_N^{\lfloor Nt\rfloor+1}(x,y)$, where $P_N^0=I$. The result follows. ■

11.7 EXERCISE. Suppose c_N and B_N are sequences of numbers and $(n\times n)$-matrices such that $c_N\to c$ and $B_N\to B$ as $N\to\infty$. Show that $\sum_{k=0}^\infty c_N B_N^k/k!\to ce^B$ for e^B defined as the $(n\times n)$ matrix $e^B=\sum_{k=0}^\infty B^k/k!$, and where the convergence is coordinatewise or in any matrix norm.

The matrix A is the generator of a Markov process $(Y_t\colon t\geq 0)$ with state space the partitions of the set $\{1,\ldots,n\}$ and transition semigroup $P_t=e^{tA}$, i.e.

$$P(Y_{s+t}=y|\,Y_s=x)=P_t(x,y).$$

This process is known as *Kingman's coalescent process.* The initial variable Y_0 is equal to the partition $\{1,\ldots,n\}=\cup_{i=1}^n\{i\}$ in one-point sets.

From the general theory of Markov processes it follows that the evolution of Y can also be described as follows. At time 0 there are n lineages forming $\binom{n}{2}$ pairs. These pairs are competing to create the first coalescing event. "Competing" means that every of the pairs generates, independently from the other pairs, a standard exponential random variable. The pair with the smallest variable wins, and the two lineages coalesce to one. There are now $n-1$ lineages, forming $\binom{n-1}{2}$ pairs; history repeats itself with this reduced set, and independently of the past. This process continues until there is only one lineage left, which is the absorbing state of the chain.

If there are still j lineages alive, then there are $\binom{j}{2}$ pairs competing and a coalescent event occurs at the minimum of $\binom{j}{2}$ independent standard exponential variables. The distribution of this minimum is exponential with mean $1/\binom{j}{2}$. Therefore, in the beginning, when there are still many lineages, the next coalescent event arrives quickly, but the mean inter arrival time steadily increases if time goes on. Figure 11.3 shows some typical realizations of the coalescent process Y, clearly illustrating that most of the coalescing events occur early.

The time T_j to go from j lineages to $j-1$ lineages is an exponential variable with intensity $\binom{j}{2}$. Therefore the expectation of the total height of the tree is

$$\mathrm{E}\sum_{j=2}^n T_j=\sum_{j=2}^n\frac{1}{\binom{j}{2}}=2\Bigl(1-\frac{1}{n}\Bigr).$$

The expected length of the time it takes the final pair of lineages to combine into one is equal to $\mathrm{E}T_2=1$ and hence is almost half the time for the total population to coalesce, if n is large.

Figure 11.3 gives a misleading picture of the coalescent process in that the individuals (on the horizontal axis) have been placed so that the random walks

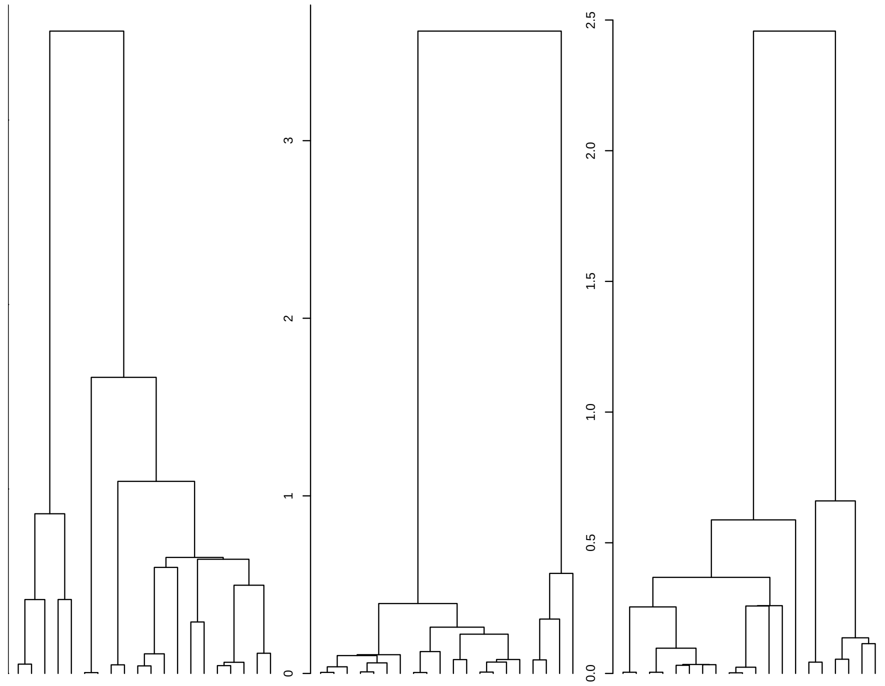

Figure 11.3. Three realizations of the n-coalescent process for $n = 20$. The individuals $1, 2, \ldots, n$ have been placed on the horizontal axis in an order so that their coalescent graph has no intersections.

describing their history do not intersect. Because typically we are not interested in the ordering of the individuals, nothing important is lost. We could also define an abstract *tree structure* by defining two sample paths of the coalescent (i.e. two sequences of nested partitions of $\{1, 2, \ldots, n\}$) to be equivalent if there is a permutation σ of the individuals (the set $\{1, 2, \ldots, n\}$) such that the second sequence of partitions applied to the individuals $\{\sigma(1), \ldots, \sigma(n)\}$ is the same as the first sequence. The coalescent process induces a probability measure on the collection of all equivalence classes for this relation.

The time unit of the coalescent process (the vertical scale in Figure 11.3) is by construction confounded with population size. In the approximating process $Y^N = (Y_t^N : t \geq 0)$ the successive generations are placed on a grid with mesh width $1/N$. Thus one time unit in the coalescent corresponds to N generations.

The coalescent process derived in the preceding theorem concerns a sample of n individuals from a population of $N \to \infty$ individuals. By construction this n-coalescent is "contained" in an $n+1$-coalescent, describing the same n individuals and another individual. Actually, the n-coalescents for all $n \in \mathbb{N}$ can be constructed consistently on a single probability space.[‡]

11.2 Robustness

Kingman's coalescent process can be viewed as an approximation to the Wright-Fisher model, but it also arises from a variety of other finite population models. In particular, suppose that the numbers $M = (M_1, \ldots, M_N)$ of offspring of individuals $1, \ldots, N$ is an exchangeable random vector such that $\sum_{i=1}^N M_i = N$. If M is not multinomial, then the rule that "children choose their parents independently at random" is not valid any more. However, the basic transition probabilities for the coalescents of the random walks connecting the generations do not change much.

Denote a random permutation of the N children again by $a_1, \ldots, a_N$.

11.8 Lemma. *The probability that k (different) children $a_{i_1}, \ldots, a_{i_k}$ choose k different parents is equal to $\mathrm{E}M_1 \cdots M_k$. The probability that children a_{i_1}, a_{i_2} choose the same parent and children $a_{i_3}, \ldots a_{i_k}$ choose $k-2$ other, different parents is equal to $(N-k+1)^{-1}\mathrm{E}M_1(M_1-1)M_2 \cdots M_{k-1}$.*

Proof. If $M = m$ for given $m = (m_1, \ldots, m_N)$, then we can represent the N children by m_1 times the symbol 1, m_2 times the symbol 2, etc. The event E that the children $a_{i_1}, \ldots, a_{i_k}$ choose different parents occurs if a random permutation of these symbols has different symbols at the positions $i_1, \ldots, i_k$. There are $N!$ permutations of the N symbols. The k different symbols could be any sequence $(j_1, \ldots, j_k)$ of indices $1 \le j_1 \ne \ldots \ne j_k \le N$ such that $m_{j_i} \ge 1$ for every i. For a given set of indices, and still given $M = m$, there are $m_{j_1} \cdots m_{j_k}$ possible choices for the indices. This number is correct even if $m_{j_i} = 0$ for some i. The remaining symbols can be ordered in $(N-k)!$ ways. It follows that

$$P(E|\, M = m) = \sum \cdots \sum_{1 \le j_1 \ne \ldots \ne j_k \le N} \frac{m_{j_1} \cdots m_{j_k} (N-k)!}{N!}.$$

We obtain the probability $P(E)$ by multiplying this by $P(M = m)$ and summing over all possible values of m. By exchanging the two sums, this can be written as

$$\sum \cdots \sum_{1 \le j_1 \ne \ldots \ne j_k \le N} \frac{(N-k)!}{N!} \mathrm{E}M_{j_1} \cdots M_{j_k}.$$

The expectation $\mathrm{E}M_{j_1} \cdots M_{j_k}$ is the same for every choice of $j_1, \ldots, j_k$, by the exchangeability of M. The display reduces to $\mathrm{E}M_1 \cdots M_k$.

With the same notation as before, the event E that the children a_{i_1}, a_{i_2} choose the same parent and children $a_{i_3}, \ldots a_{i_k}$ choose $k-2$ other, different parents occurs if a random permutation of m_1 times the symbol 1, m_2 times the symbol 2,

[‡] See Kingman ??

etc. has the same symbol at positions i_1 and i_2 and different symbols at positions $i_3,\ldots,i_k$. This involves $k-1$ different symbols $j_1,\ldots,j_{k-1}$, which we can assign to positions $i_1,\ldots,i_k$ in the form $j_1,j_1,j_2,\ldots,j_{k-1}$. By the same arguments as before, the conditional probability of E is therefore

$$P(E|\, M=m)=\sum_{1\le j_1\ne\ldots\ne j_{k-1}\le N}\!\!\cdots\sum \frac{m_{j_1}(m_{j_1}-1)m_{j_2}\cdots m_{j_{k-1}}(N-k)!}{N!}.$$

This readily leads to the unconditional probability claimed in the lemma. ∎

11.9 EXERCISE. Show that the probability that different children $a_{i_1},\ldots,a_{i_k}$ choose parents $j_1,\ldots,j_k$ is equal to $1/(N)_k\,\mathrm{E}\prod_{i=1}^N(M_i)_{n_i}$, for $n_i=\#(r\colon j_r=i)$ and $(N)_k=N(N-1)\cdots(N-k+1)$ for any natural numbers N and $k\le N$. [The lemma is the special case that $(n_1,\ldots,n_N)$ is a vector with coordinates 0 or 1 only, or with one coordinate 2 and other coordinates 0 or 1.]

The distribution (and in fact even the dimension) of the vector M depends on N, which we would like to send to infinity. For clarity write M^N instead of M. By exchangeability and the fact that $\sum_{i=1}^N M_i^N=N$ the first marginal moment satisfies $\mathrm{E}M_1^N=1$ for every N. Under the condition that the marginal variance tends to a limit and the third order marginal moments are bounded in N, we can expand the expectations in the preceding lemma in powers of $1/N$.

11.10 Lemma. *If* $\operatorname{var}M_1^N\to\sigma^2$ *and* $\mathrm{E}(M_1^N)^3=O(1)$ *as* $N\to\infty$, *then*

$$\mathrm{E}M_1^N\cdots M_k^N=1-\binom{k}{2}\frac{\sigma^2}{N}+o\Bigl(\frac{1}{N}\Bigr),$$

$$\frac{1}{N-k+1}\mathrm{E}M_1^N(M_1^N-1)M_2^N\cdots M_{k-1}^N=\frac{\sigma^2}{N}+o\Bigl(\frac{1}{N}\Bigr).$$

Proof. Omitting the superscript N from the notation, and writing $M_i=1+(M_i-1)$, we can expand

$$\begin{aligned}&\mathrm{E}M_1\cdots M_k\\&=1+\mathrm{E}\sum_{i=1}^k(M_i-1)+\mathrm{E}\sum\sum_{1\le i<j\le k}(M_i-1)(M_j-1)+\cdots+\mathrm{E}\prod_{i=1}^k(M_i-1).\end{aligned}$$

Here the second term on the right vanishes, because $\mathrm{E}M_i=1$ for every i, by exchangeability. For the first assertion of the lemma it suffices to show that

(i) $\mathrm{E}(M_1-1)(M_2-1)=-\sigma^2/(N-1)$.

(ii) $\mathrm{E}(M_1-1)\cdots(M_k-1)=o(1/N)$ for $k\ge 3$.

For any $k\ge 2$ we can write, by exchangeability,

$$\mathrm{E}(M_1-1)\cdots(M_k-1)=\frac{1}{N-k+1}\sum_{j=k}^N\mathrm{E}(M_1-1)\cdots(M_{k-1}-1)(M_j-1).$$

Because $\sum_{j=1}^N M_j = N$ by assumption, we can replace $\sum_{j=k}^N (M_j - 1)$ by $-\sum_{j=1}^{k-1}(M_j - 1)$, and next simplify the resulting expression to, again using exchangeability,

$$-\frac{k-1}{N-k+1}\mathrm{E}(M_1-1)^2(M_2-1)\cdots(M_{k-1}-1).$$

For $k = 2$ this yields assertion (i). To prove assertion (ii) we repeat the argument, and rewrite the preceding display as

$$\begin{aligned}&\frac{k-1}{N-k+1}\frac{1}{N-k+2}\mathrm{E}(M_1-1)^2(M_2-1)\cdots(M_{k-2}-1)\sum_{j=1}^{k-2}(M_j-1)\\ &\quad=\frac{k-1}{N-k+1}\frac{1}{N-k+2}\Bigl[(k-3)\mathrm{E}(M_1-1)^2(M_2-1)^2(M_3-1)\cdots(M_{k-2}-1)\\ &\qquad\qquad\qquad+\mathrm{E}(M_1-1)^3(M_2-1)\cdots(M_{k-2}-1)\Bigr].\end{aligned}$$

To prove (ii) it suffices to show that the two expectations appearing in the brackets on the right are of order $o(N)$, for $k \geq 3$.

For any integers $p, q \geq 0$ and $k \geq 3$ we have

$$\begin{aligned}&\mathrm{E}|M_1-1|^p|M_2-1|^q|M_3-1|\cdots|M_k-1|\\ &\quad=\frac{1}{N-k+1}\sum_{j=k}^N \mathrm{E}|M_1-1|^p|M_2-1|^q|M_3-1|\cdots|M_{k-1}-1||M_j-1|\\ &\quad\leq\frac{2N}{N-k+1}\mathrm{E}|M_1-1|^p|M_2-1|^q|M_3-1|\cdots|M_{k-1}-1|,\end{aligned}$$

because $\sum_{j=k}^N |M_j - 1| \leq 2N$. By induction we see that the preceding display is bounded by

$$\frac{2N}{N-k+1}\cdots\frac{2N}{N-2}\mathrm{E}|M_1-1|^p|M_2-1|^q.$$

For $p = 3$ and $q = 0$ this proves that $\mathrm{E}(M_1 - 1)^3(M_3 - 1)\cdots(M_k - 1)$ is bounded in N for $k \geq 3$, and hence certainly $o(N)$. For $p = 2$ and $q = 2$ we argue further that $\sum_j (M_j - 1)^2 = \sum_j M_j^2 - N \leq N(\max_j M_j - 1)$ and hence

$$\begin{aligned}\mathrm{E}(M_1-1)^2(M_2-1)^2&=\frac{1}{N-1}\mathrm{E}(M_1-1)^2\sum_{j=2}^N(M_j-1)^2\\ &\leq\frac{N}{N-1}\mathrm{E}(M_1-1)^2\max_j|M_j-1|\\ &\leq\frac{N}{N-1}\bigl(\mathrm{E}|M_1-1|^3\bigr)^{2/3}\bigl(\mathrm{E}\max_j|M_j-1|^3\bigr)^{1/3}\\ &\leq\frac{N}{N-1}\mathrm{E}|M_1-1|^3N^{1/3}.\end{aligned}$$

In the last step we use that a maximum of nonnegative variables is smaller than the sum. Again the upper bound is $o(N)$. The proof of (ii) is complete.

For the second assertion of the lemma we expand

$$\mathrm{E}M_1(M_1-1)M_2\cdots M_k = \mathrm{E}(M_1-1)\times$$
$$\times\Bigl[1+\sum_{i=1}^k (M_i-1)+\sum\sum_{1\le i<j\le k}(M_i-1)(M_j-1)+\cdots+\prod_{i=1}^k (M_i-1)\Bigr].$$

Because $\mathrm{E}(M_1-1)=0$ and $\mathrm{E}(M_1-1)^2\to\sigma^2$, it suffices to show that in addition to (ii) as given previously, also
(iii) $\mathrm{E}(M_1-1)^2(M_2-1)\cdots(M_k-1)=o(1)$ for $k\ge 2$.
This has already been established in the course of the proof of (ii). ■

Combining the two lemmas we see that again the probability of more than two children choosing the same parent or two of more sets of children choosing the same parent are of the lower order $O(1/N^2)$. Following the approach of Section 11.1, we obtain the same continuous time approximation, with the only difference that the generator A of the limiting Markov process is multiplied by the variance parameter σ^2:

$$A(x,y)=\begin{cases} -\sigma^2\binom{\#x}{2}, & \text{if } y=x,\\ \sigma^2, & \text{if } x\Rightarrow y,\\ 0, & \text{otherwise.}\end{cases}$$

Multiplying the generator with a constant is equivalent to linearly changing the time scale: the Markov proces with the generator in the display is the process $t\mapsto Y_{\sigma^2 t}$ for Y the standard coalescent process. Thus if $\sigma^2>1$ the generations coalesce faster than for the standard coalescent process. It is said that the *effective generation size* is presently equal to N/σ^2. Of course, this only makes sense when comparing to the standard Wright-Fisher model.

In the present model with $\sigma^2>1$ the parents have a more varied number of offspring, where the variation is mostly due to occasional larger offspring, the mean offspring number still being 1. This explains that we need to go back fewer generations to find a common ancestor.

11.3 Varying Population Size

In the Wright-Fisher model the successive populations are assumed to have the same size. For many applications this is not realistic. In this section we show that evolution with varying population can also be approximated by Kingman's coalescent process, but with a rescaled time. The intuition is that, given a smaller population of possible parents, children are more likely to choose the same parent, thus leading to more rapid coalescence. With fluctuating population sizes the time scale for coalescence would shrink or extend proportionally.

We describe the evolution in backward time. Suppose that the population at time k consists of N_k individuals, and that the N_{k-1} individuals at time $k-1$ choose their parents from the N_k individuals at time k at random and independently of each other. At time 0 we start N_0 random walks, consisting of children choosing their parents. At each time we define X_k to be the partition of the set $\{1,2,\ldots,N_0\}$ corresponding to the random walks that have coalesced by that time. If $X_{k-1}=x$, then there are $\#x$ separate walks at time $k-1$. The probabilities that these walks do not coalesce or that exactly one pair of walks coalesces in the time interval between $k-1$ and k can be calculated as before, the only difference being that the $\#x$ children now have a number N_k of parents that varies with k to choose from. It follows that, for any partitions x and y with $x \Rightarrow y$,

$$P(X_k = x|\, X_{k-1}=x) = 1-\binom{\#x}{2}\frac{1}{N_k}+O\Big(\frac{1}{N_k^2}\Big),$$
$$P(X_k = y|\, X_{k-1}=x) = \frac{1}{N_k}+O\Big(\frac{1}{N_k^2}\Big).$$

The remainder terms $R_{N,k}=O(1/N_k^2)$ have the property that $N_k^2 R_{N,k}$ remains bounded if $N_k\to\infty$.

To study the limiting situation we focus again on the walks originating from a fixed finite set of individuals at (backward) time 0, and consider the partitions induced on this set of walks. We suppose that all population sizes N_k depend on a parameter N such that $N_k\to\infty$ for every k as $N\to\infty$, and let $X^{n,N}=(X_k^{n,N}: k\in\mathbb{N})$ be the corresponding Markov chain on the collection of partitions of the set $\{1,\ldots,n\}$. By the preceding paragraph the chain $X^{n,N}$ possesses the transition matrix $\mathcal{P}_{N,k}$ at (backward) time k (giving the probabilities $\mathcal{P}_{N,k}(x,y)=P(X_k^{n,N_k}=y|\,X_{k-1}^{n,N_k}=x)$) such that $N_k(\mathcal{P}_{N,k}-I)\to A$, for A the matrix given in (11.5). Define a grid of time points

$$t_0^N=0,\qquad t_k^N=\sum_{i=1}^k\frac{1}{N_i},\quad k\ge 1,$$

and define a continuous time stochastic process $(Y_t^N: t\ge 0)$ by

$$Y_{t_k^N}^N = X_k^{n,N},\qquad k=0,1,2,\ldots,$$

and define Y_t^N to be constant on the intervals $[t_k^N, t_{k+1}^N)$.

11.11 Theorem. *[ASSUME SOME REGULARITY??] The transition matrices $Q_t^{(N)}$ defined by $Q_t^{(N)}(x,y)=P(Y_{s+t}^N=y|\,Y_s^N=x)$ satisfy $Q_t^{(N)}\to e^{tA}$ as $N\to\infty$, for any $t>0$.*

Proof. The transitions of the process Y^N in the interval $(s,s+t]$ are the transitions of the process $X^{n,N}$ at the time points $k+1,\ldots,l$ such that $t_k^N\le s<t_{k+1}^N\le t_l^N\le$

$s+t<t_{l+1}^N$. The corresponding transition matrix is

$$\prod_{i=k+1}^{l}\mathcal{P}_{N,i}=\prod_{i=k+1}^{l}\Bigl(I+\frac{1}{N_i}N_i(\mathcal{P}_{N,i}-I)\Bigr).$$

The terms in the product are approximately equal to $I+N_i^{-1}A$. The product is asymptotically equivalent to (need some regularity??)

$$\prod_{i=k+1}^{l}e^{N_i^{-1}A}=e^{(t_l^N-t_k^N)A}.$$

The right side tends to e^{tA} as $N\to\infty$. ∎

Thus again the discrete time process can be approximated by the coalescent process. In order to obtain the standard coalescent as an approximation it was necessary to place the generations in a nonuniform way on the time axis. For instance, an exponentially increasing population size corresponds to the scheme $N_{k-1}=\alpha N_k$ (with $\alpha>1$), yielding $N_k=\alpha^{-k}N_0$ by iteration, and hence the time points

$$t_k^N=\sum_{j=1}^{k}\frac{\alpha^j}{N_0}=C(\alpha^k-1),\qquad C=\frac{\alpha}{N_0(\alpha-1)}.$$

Thus the population at backward generation k is represented by $Y_{C(\alpha^k-1)}$ for $(Y_t\colon t\ge 0)$ the standard coalescent. [DOES EXPONENTIAL GROWTH SATISFY REGULARITY?]

In general, the standard coalescent gives the correct ordering of the coalescing events, but on an unrealistic time-scale. A process of the form $t\mapsto Y_{g(t)}$ for a $g\colon[0,\infty)$ a monotone transformation and $(Y_t\colon t\ge 0)$ the standard coalescent is the correct approximation to the population process in real time.

11.4 Diploid Populations

Human (and in fact most other) procreation is sexual and involves pairs of parents rather than single haplotypes, as in the Wright-Fisher model. This complicates the coalescence process, but Kingman's continuous time process still arises as an approximation. We consider single locus haplotypes, thus still ignoring recombination events.

The "individuals" in our successive populations will also still be alleles (haplotypes), not persons or gene pairs. Thus coalescence of two lineages (of alleles) corresponds to the members of the lineages being IBD with the parent allele in which they come together. The tree structure we are after has alleles as its nodes, and is not like a family pedigree, in which persons (allele pairs) are the nodes.

The new feature is that the haplotypes are organized in three ways: they form gene pairs, a gene pair may be male or female, and within each gene pair one haplotype is paternal and one is maternal. Suppose that each generation consists of N male gene pairs and N female gene pairs, thus $4N$ alleles in total. A new generation is formed by

- Pairing the N males and N females at random ("random mating").
- Each couple has S_j sons and D_j daughters, the vectors $(S_1, \ldots, S_N)$ and $(D_1, \ldots, D_N)$ are independent and possess multinomial distributions with parameters $(N, 1/N, \ldots, 1/N)$ ("Wright-Fisher offspring").
- Each parent segregates a random allele of his or her pair of alleles to each offspring, independently across offspring ("Mendelian segregation").

This scheme produces successive generations of N males (sons) and N females (daughters), who have $4N$ alleles in total, $2N$ of which are paternal (originate from a male) and $2N$ are maternal (originate from a female). Each of the $4N$ alleles in a generation is a copy of an allele in the preceding generation, and by following this up in history any allele can be traced back to an allele in the first generation. We are interested in the coalescence of the paths of the alleles, as before.

11.5 Mutation

Because coalescence happens with probability one if we go far enough back into the past, all present day alleles are copies of a single founder allele. If there were no mutations, then all present day individuals would be homozygous and identical. Mutations divide them in a number of different types.

Mutations of the alleles are usually superimposed on the coalescent tree as an independent process. In the *infinite alleles model* every mutation leads to a new allele, which is then copied exactly to the offspring until a new mutation arises, which creates a completely new type. (We are now using the word "allele" in its meaning of a variant of a gene.) Motivation for such a model stems from viewing a locus as a long sequence of nucleotides and assuming that each mutation concerns (changes, inserts or deletes) a single nucleotide. In view of the large number of nucleotides, which can each mutate to three other nucleotides, it is unlikely that two sequences of mutations would yield the same result or lead back to the original.

In the continuous time approximation mutations are assumed to occur along the branches of the coalescent tree according to Poisson processes, where each branch of the coalescent three has its own Poisson process and the Poisson processes along different branches are assumed independent. Because the Poisson process arises as the continuous time limit of an events process consisting of independent Bernoulli numbers of events in disjoint intervals of decreasing length, this approximation corresponds to mutations in the discrete-time Wright-Fisher model that occur with probability of the order $O(1/N)$ in each generation and independently across generations and offspring.

We may now study the number of different types of alleles in a given set of n individuals. The distribution of the number of types turns out to be the same as in a simple model involving coloured marbles, called *Hoppe's urn model*, which we now describe. At time k an urn contains one black marble and k other marbles of various colours, where some colours may appears multiple times. At time $k = 1$ the urn only contains the black marble. The black marble has weight $\theta > 0$, whereas each coloured marble has weight one. We choose a marble from the urn with probability proportional to its weight. If the marble is black, then we put it back in the urn and add a marble of a colour that was not present yet. If the marble is coloured, then we put it back in the urn and add a marble of the same colour. The urn now contains $k + 1$ marbles, and we repeat the experiment. At time n the urn contains n coloured marbles, and we may define random variables $A_1, \ldots, A_n$ by

$$A_i = \# \text{ of colours that appear } i \text{ times in the urn.}$$

Obviously $\sum_{i=1}^{n} iA_i$ is the total number of marbles in the urn and hence $\sum_{i=1}^{n} iA_i = n$. It turns out that the vector $(A_1, \ldots, A_n)$ also gives the number of different type of individuals in the coalescent model with mutation. The simplicity of Hoppe's urn model helps to compute the distribution of this vector.

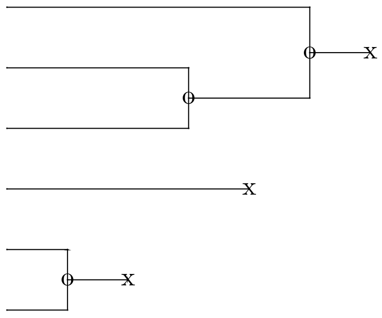

Figure 11.4. Realization from Hoppe's urn model. Crosses "x" and circles "o" indicate that the black marble or a coloured marble is drawn. Drawing starts on the right with the black marble and proceeds to the left. Each line indicates a marble in the urn, and splitting of a line indicates an additional marble of that colour. The events are separated by 1 time unit (horizontal axis).

That Hoppe's model gives the right distribution is best seen from a graphical display of the realizations from the urn model, as in Figure 11.4. Time is passing from right to left in the picture, with the cross on the far right indicating the black marble being drawn at time 0, leading to a first coloured marble in the urn. Circles and crosses to the left indicate coloured marbles or the black marble being drawn. Each path represents a coloured marble.

At first sight the Hoppe graph has nothing in common with the coalescent graph. One reason is that the time intervals between "events" in Hoppe's model are fixed to unity, whereas the intervals in Kingman's coalescent are exponentially distributed. This difference influences the horizontal scale, but is irrelevant for the distribution of the types. A second difference is that the Hoppe graph is disconnected, whereas in the coalescent graph each pair of random walks comes together at some point. This difference expresses the different types of individuals caused by mutations. In the infinite alleles model two individuals are of the same type if and only if their random walks coalesce at their MCRA without a mutation event occurring on either of the two paths to the MCRA. If we slide from left to right in the coalescent graph, then whenever we meet a mutation event on a given path, the individuals whose random walks coalesce in this path are of different type than the other individuals. The coalescent graph could be adapted to show the different types by removing the paths extending to the right from such a mutation event. This is called *killing* of the coalescing random walks after mutation. After killing all the paths in this way, moving from left to right, backwards in time, the coalescent graph has the same structure as Hoppe's graph.

The preceding describes the qualitative relationship between the different types in the coalescent and Hoppe's model. Quantitatively these models agree also. If k walks in the coalescent graph have not been killed, then the next event is a coalescence of one of the $\binom{k}{2}$ pairs of walks or a mutation on one of the k open paths. Because the events are inserted according to independent Poisson processes ($\binom{k}{2}$ with intensity 1 and k with intensity μ), the relative probabilities of the two types of events (coalescence or mutation) are $\binom{k}{2}$ and $k\mu$, respectively. If there are k paths open in the (backward) Hoppe graph, then the next time to the left corresponds to drawing from the urn when it contains $k-1$ coloured marbles and 1 black marble. With relative probabilities $k-1$ and θ a coloured marble is drawn, yielding a circle and a coalescence in the graph, or the black marble, yielding a cross. In both models the coalescing events occur equally likely on each path. The quotients of the relative probabilities of coalescence or killing in the two models are

$$\frac{\binom{k}{2}}{k\mu} = \frac{k-1}{2\mu}, \qquad \text{and} \qquad \frac{k-1}{\theta}.$$

Thus the two models correspond if $\theta = 2\mu$. It is relatively straightforward to compute a number of probabilities of interest in Hoppe's model.

Let $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$.

11.12 Theorem (Ewens' sampling formula). *Let $A_{n,i}$ be the number of types present i times in the n-coalescent with mutation at rate $\theta/2$ per branch. Then for any $(a_1 \ldots, a_n)$ with $\sum_{i=1}^n i a_i = n$,*

$$P(A_{n,1} = a_1, \ldots, A_{n,n} = a_n) = \frac{n!}{\theta_{(n)}} \prod_{i=1}^{n}{}^{n} \frac{(\theta/i)^{a_i}}{a_i!}.$$

Proof. Using induction on n, we prove that the number of types present at time n in Hoppe's urn model satisfies the equation. For $n = 1$ necessarily $a_1 = 1$ and the assertion reduces to $P(A_{1,1} = 1) = 1$, which is correct.

Assume that the assertion is correct for given $n \geq 1$, and consider the assertion for $n + 1$. If we define $A_{k,i} = 0$ for $i > k$, then the vectors $(A_{k,1}, \ldots, A_{k,n+1})$ for $k = 1, 2, \ldots$ have the same dimension and form a Markov chain, with two possible types of transitions at each time step:

(i) The black ball is drawn. At time n this has probability $\theta/(\theta + n)$ and yields a transition from $(a_1, a_2, \ldots, a_n, 0)$ to $(a_1 + 1, a_2, \ldots, a_n, 0)$.

(ii) A coloured ball is drawn of which i balls are present in the urn. At time n this has probability $i/(\theta + n)$ and yields a transition from $(a_1, a_2, \ldots, a_n, 0)$ to $(a_1, a_2, \ldots, a_i - 1, a_{i+1} + 1, \ldots, a_n, 0)$. There are a_i coloured balls that all yield this same transition, so that the total probability of this transition is equal to $ia_i/(\theta + n)$.

Let E_0 and E_i be the events of the types as in (i) and (ii).

If $a_{n+1} = 0$, then

$$\begin{aligned}
&P\left(A_{n+1,1} = a_1, \ldots, A_{n+1,n+1} = a_{n+1}\right)\\
&= P\left(E_0, A_{n,1} = a_1 - 1, A_{n,2} = a_2 \ldots, A_{n,n} = a_n\right)\\
&+ \quad \sum_{j=1}^{n} P\left(E_j, A_{n,1} = a_1, \ldots, A_{n,j} = a_j + 1, A_{n,j+1} = a_{j+1} - 1, \ldots A_{n,n} = a_n\right).
\end{aligned}$$

In view of the induction hypothesis we can rewrite this as

$$\frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!} \Bigl[\frac{\theta}{\theta + n}\frac{a_1}{\theta/1} + \sum_{j=1}^{n} \frac{j(a_j + 1)}{\theta + n}\frac{a_j}{\theta/j}\frac{\theta/(j + 1)}{a_{j+1}}\Bigr].$$

The identity $a_1 + \sum_{j=1}^{n}(j + 1)a_{j+1} = n + 1$ permits to write this in the desired form.

If $a_{n+1} = 1$, whence $a_1 = \cdots = a_n = 0$, then only a transition of type (ii) can have occurred, and we can write the probability of the event of interest as

$$\begin{aligned}
&\frac{n}{\theta + n} P(A_{n,1} = 0, \ldots, A_{n,n-1} = 0, A_{n,n} = 1)\\
&\qquad = \frac{n}{\theta + n}\frac{n!}{\theta_{(n)}}\frac{\theta}{n} = \frac{(n + 1)!}{\theta_{(n+1)}}\frac{\theta}{n + 1}.
\end{aligned}$$

This concludes the proof. ∎

11.13 Theorem. *The number of different types K_n in a sample of n individuals satisfies $\mathrm{E}K_n \sim \theta \log n$ and $\operatorname{var} K_n \sim \theta \log n$, as $n \to \infty$. Moreover the sequence $(K_n - \mathrm{E}K_n)/\operatorname{sd}(K_n)$ converges in distribution to the standard normal distribution.*

Proof. The variable K_n is equal to the number of times the black marble was drawn in the first n draws from Hoppe's urn. It can be written as $K_n = \sum_{i=1}^{n} \Delta_i$

for Δ_i equal to 1 if the ith draw yielded the black marble and 0 otherwise. Because $P(\Delta_i = 1) = \theta/(\theta + i)$, it follows that

$$\mathrm{E}K_n = \sum_{i=1}^n \mathrm{E}\Delta_i = \sum_{i=1}^n \frac{\theta}{\theta + i},$$
$$\operatorname{var} K_n = \sum_{i=1}^n \operatorname{var} \Delta_i = \sum_{i=1}^n \frac{\theta}{\theta + i} - \sum_{i=1}^n \Bigl(\frac{\theta}{\theta + i}\Bigr)^2.$$

The inequalities

$$\int_1^n \frac{\theta}{\theta + x}\,dx \leq \sum_{i=1}^n \frac{\theta}{\theta + i} \leq \int_0^n \frac{\theta}{\theta + x}\,dx$$

yield that

$$\frac{\theta\bigl(\log(\theta + n) - \log(\theta + 1)\bigr)}{\theta \log n} \leq \frac{\mathrm{E}K_n}{\theta \log n} \leq \frac{\theta\bigl(\log(\theta + n) - \log \theta\bigr)}{\theta \log n}.$$

Here left and right tend to 1 as $n \to \infty$. The second sum in the expression for $\operatorname{var} K_n$ is bounded by $\sum_{i=1}^\infty \theta^2/(\theta + i)^2 < \infty$ and hence is negligible relative to the first sum, which tends to infinity.

The asymptotic normality is a consequence of the Lindeberg-Feller central limit theorem. ■

11.6 Recombination

[♭] If we are interested in the ancestry of haplotypes that can undergo recombination during meiosis, then the basic coalescent model is insufficient, as a multi-locus haplotype can have different parents for the various loci. (Here "parent" is understood to be a haplotype, or chromosome, not a diploid individual.) A simple way around this would be to construct a separate coalescent tree for every locus. Due to recombination these trees will not be the same, but they will be related for loci that are not too distant. In this section it is shown that the trees corresponding to different loci can be incorporated in a single graph, called the *recombination graph.* Perhaps a little bit surprising is that there is a single individual at the root of this graph, who is an ancestor for the sample at all the loci.

The main structure already arises for haplotypes consisting of two loci. Consider a population of N individuals, each consisting of two linked loci, referred to as "L" and "R", as shown in Figure 11.5. For simplicity we consider the individuals as haploid and let reproduction be asexual. As before it is easiest to describe the

[♭] CHECK SECTION FOR SERIOUS ERRORS??

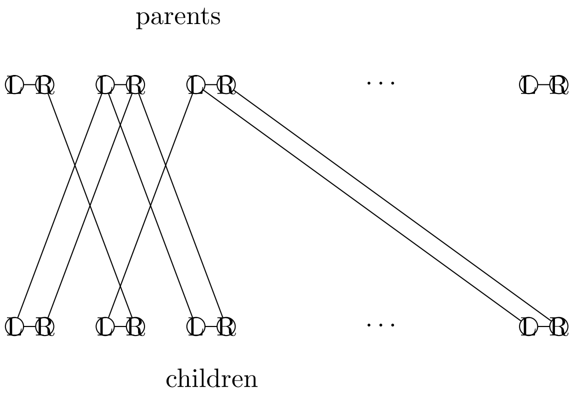

Figure 11.5. A population of two-loci haploid parents and their children.

relation between the population of N two-loci parents and N two-loci offspring backwards in time:

- With probability $1 - r$ a child chooses a single parent and copies both loci from the parent.
- With probability r a child chooses two different parents and copies the "L"-locus from the first and the "R"-locus from the second parent.
- The children choose independently of each other.

The parameter r is the recombination fraction between the two loci.

Given this reproduction scheme we follow the ancestry of a fixed number $n \ll N$ of two-loci individuals backwards in time, where we keep track of *all* parents that provide genetic material, whether they pass on material on a single locus or for both loci. Thus it is possible that the number of lines increases (if one or more children recombine the "L" and "R" loci from two parents) as well as decreases (if two children do not recombine and choose the same parent). Before studying the resulting partitioning structure we consider the total number of ancestors (lines) as a process in time. Let $A_0^N = n$ and, for $j = 1, 2, \ldots$, let A_j^N be the number of ancestors of the genetic material of the n individuals. We shall show that, asymptotically as the population size N tends to infinity and under the assumption that the recombination fraction r tends to zero, the process A^N is a birth-death process that will reach the state 1 eventually. This shows that if we trace back far enough in the past there is an individual who is the ancestor of the n individuals for *both* loci.

We assume that the recombination fraction takes the form $r = \rho/(2N)$ for a positive constant ρ. The crux is that for $N \to \infty$ only three possible events contribute significantly to the ancestry process. Given a number of k children in a given generation, these are the events:

NC All k children choose a different parent without recombination ("no change").

R Exactly one child chooses two parents and recombines, and the other $k - 1$ children choose a different parent without recombination ("recombination").

C Exactly two children choose the same parent, and the other $k-2$ children choose a different parent, all of them without recombination ("coalescence").

The first event, which turns out to account for most of the probability, makes no change to A^N, whereas the second and third events cause an increase or decrease by 1 respectively. This observation leads to the inequalities:

$$\begin{aligned}
P(A^N_{j+1}=k|\,A^N_j=k)\geq P(NC)&=\Bigl(1-\frac{\rho}{2N}\Bigr)^k\frac{N(N-1)\cdots(N-k+1)}{N^k}\\
&=1-\frac{\rho}{2N}-\frac{\binom{k}{2}}{N}+O\Bigl(\frac{1}{N^2}\Bigr),\\
P(A^N_{j+1}=k+1|\,A^N_j=k)\geq P(R)&=\binom{k}{1}\Bigl(1-\frac{\rho}{2N}\Bigr)^{k-1}\frac{N(N-1)\cdots(N-k)}{N^{k+1}}\\
&=\frac{k\rho}{2N}+O\Bigl(\frac{1}{N^2}\Bigr),\qquad (11.14)\\
P(A^N_{j+1}=k-1|\,A^N_j=k)\geq P(C)&=\Bigl(1-\frac{\rho}{2N}\Bigr)^k\binom{k}{2}\frac{N(N-1)\cdots(N-k+2)}{N^k}\\
&=\frac{\binom{k}{2}}{N}+O\Bigl(\frac{1}{N^2}\Bigr).
\end{aligned}$$

The sum of the right sides of this display is equal to $1-O(1/N^2)$. Because the transition probabilities on the left sides add to a number not bigger than one, the inequalities must be equalities up to the order $O(1/N^2)$. This show that for $N\to\infty$ the three types of events account for all relevant transitions, where the "no change" transition is by far the most likely one. Define a continuous time process $B^N=(B^N_t\colon t\geq 0)$ by

$$B^N_{j/N}=A^N_j,$$

and letting B be continuous on the intervals $\bigl[j/N,(j+1)/N\bigr)$. Then the sequence of processes B^N tends to a Markov process $B=(B_t\colon t\geq 0)$ on the state space $\{1,2,3,\ldots,\}$ with generator given by

$$C(k,l)=\begin{cases}\binom{k}{2}, & \text{if } l=k-1,\\ -\frac{k\rho}{2}-\binom{k}{2}, & \text{if } k=l,\\ \frac{k\rho}{2}, & \text{if } k=l+1.\end{cases}$$

11.15 Theorem. *The transition matrices $Q^{(N)}_t$ defined by $Q^{(N)}_t(k,l)=P(B^N_{s+t}=l|\,B^N_s=k)$ satisfy $Q^{(N)}_t\to e^{tC}$ as $N\to\infty$, for any $t>0$.*

The proof is similar to the proof of Theorem 11.6. The limiting process B is a *birth-death process*, i.e. a Markov process on the natural numbers whose jumps are always an increase or a decrease by exactly one. The (total) death rate $\binom{k}{2}$ of the process if there are k lines "alive" is for large k much bigger than the (total) birth rate $k\rho/2$. This will cause the process to come down to the state 1 eventually.

At that point the ancestry process has reached a single individual whose two loci are the ancestor of the two loci of all n individuals. The first such individual is a two-loci MRCA of the sample. (The death rate of the process B in state 1 is zero, while the birth rate is positive; hence the process will bounce back up from there, but its relevance to the ancestry stops at the MRCA.)

Standard theory on birth-death processes allows to compute the mean time to reaching the MRCA and the distribution of the maximal number of individuals until this time. Let

$$T_{MRCA} = \inf\{t \geq 0\colon B_t = 1\}, \qquad M = \max\{B_t\colon 0 \leq t \varepsilon T_{MRCA}\}.$$

Then

$$E_n T_{MRCA} = \frac{2}{\rho}\int_0^1 \Bigl(\frac{1-v^{n-1}}{1-v}\Bigr)\bigl(e^{\rho(1-v)}-1\bigr)\,dv.$$

Furthermore

$$P_n(M\leq m) = \frac{\sum_{j=n-1}^{m-1} j!\rho^{-j}}{\sum_{j=0}^{m-1} j!\rho^{-j}}, \qquad m\geq n.$$

It can be seen from the latter formula that the maximal number of individuals in the ancestry process does not exceed the sample size n much.

11.16 EXERCISE. Prove that the distribution of M satisfies the recursion formula $P_n(M\leq m) = P_{n-1}(M\leq m)(n-1)/(\rho+n-1) + P_{n+1}(M\leq m)\rho/(\rho+n-1)$. Derive the the expression for $P_n(M\leq m)$ from this.

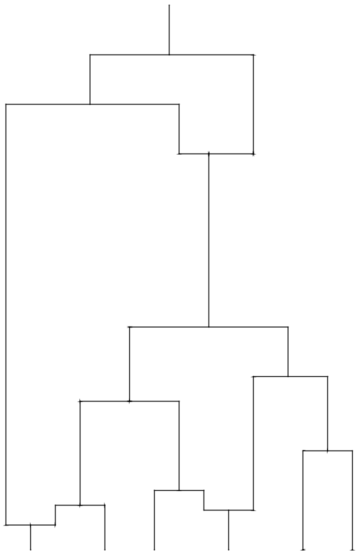

Figure 11.6. Ancestral recombination graph for a sample of 6 two-loci individuals. Calendar time flows vertically from top to bottom.

Now that we have proved that the total number of ancestors (partial or two-loci) is a birth-death process that decreases to one eventually (in the limit as $N \to \infty$), we can study the ancestral relationships in more detail. The *ancestral recombination graph* visualizes the process; see Figure 11.6 for an example. It is read backwards in time, starts with n lineages, and undergoes both coalescence and branching. A coalescence corresponds to (exactly) two children choosing the same parent (the event C) as before. The new element is the branching of a lineage, which corresponds to (exactly) one child choosing two parents (the event RC) and copying one locus from the first and the other locus from the second parent. In the ordered version of the graph (as in the figure) the lines representing the two parents are drawn with the left line representing the parent who segregates the left locus and the right line representing the parent who segregates the right locus.

Corresponding to the graph we can define a Markov process, as follows. Start with n lineages. At each given time:

- Every lineage that exists at that time generates an exponential variable with intensity $\rho/2$.
- Every pair of lineages generates an exponential variable with intensity 1.
- All exponential variables are independent.
- The lineage or pair with the smallest exponential variables wins: if it is a lineage, then this splits in two; if it is a pair, then this pair coalesces.
- The split or coalescence is inserted in the graph separated from the previous event by the winning time.

Net the process repeats, independently from the past. Eventually the process will reduce to one lineage. At that point it stops.

This process is the limit of the ancestral process described in discrete time before. The factors $\binom{k}{2}$ and k in (11.14) correspond to the number of pairs trying to coalesce and the number of lines trying to split. We omit the details of the mathematical limit procedure.

The ancestral recombination graph allows to follow the ancestorship relations for both loci and hence contains the coalescent trees for both loci. The two trees are obtained by removing at each branchpoint the paths to the right or the paths to the left, respectively. The coalescent trees for the two loci corresponding to the ancestral recombination graph in Figure 11.6 are given in Figure 11.7. Note that in this case the most recent common ancestors for the two loci, indicated by “M” in the figures, are different, and are also different from the most recent common ancestor for the two loci jointly. The latter of course is farther back in the past.

The preceding can be extended to haplotypes of more than two loci. In the model of discrete generations, a child is then still allowed to choose at most two parents, but it may divide the set of loci in any way over the two parents, thus choosing multiple crossover points. However, in the preceding setting where the probability of recombination was set to converge to zero ($r = \rho/N$ with $N \to \infty$) it is reasonable to assume that the probability of multiple crossovers is negligible relative to the probability of a single recombination. Hence in the limit approximation multiple crossovers do not occur and the left and right edges of a branching in the ancestral recombination graph can still be understood to refer to a “left”

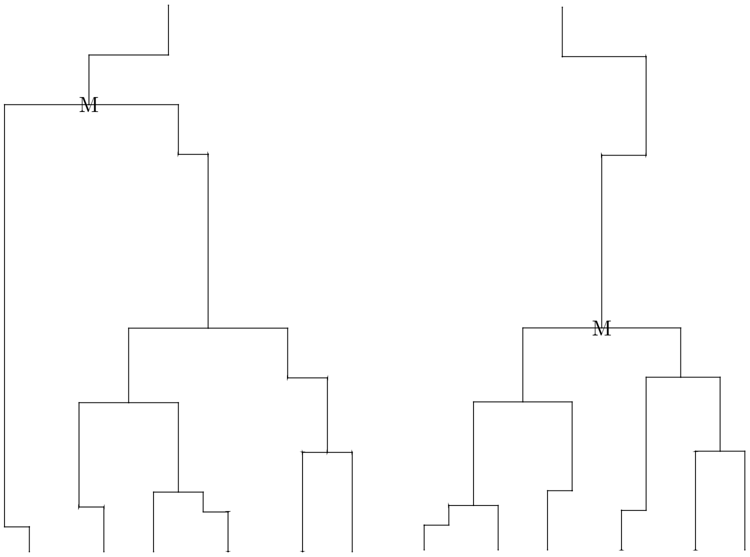

Figure 11.7. The coalescent trees of the left and right locus corresponding to the ancestral recombination graph in Figure 11.6. The bends in the paths are retained from the latter figure, to facilitate comparison. The MRCAs of the loci are indicated by the symbol "M".

and "right" arm of the genome, even though the split point between the two arms can vary. The locations of these split points are typically modelled as being independently superimposed on the ancestral recombination graph, according to a given marginal distribution, for instance the uniform distribution on the genome represented as the interval $[0, 1]$. The split points are indicated by numbers on the ancestral recombination graph, as shown in Figure 11.8. Given an annotated graph of this type, it is possible to recover the coalescent tree for any given locus x by following at each branching the appropriate route: if the branching is annoted by the number s, then left if $x < s$ and right if $x > s$.

Notes

Ewens, Kingman, Moehle, Sagitov

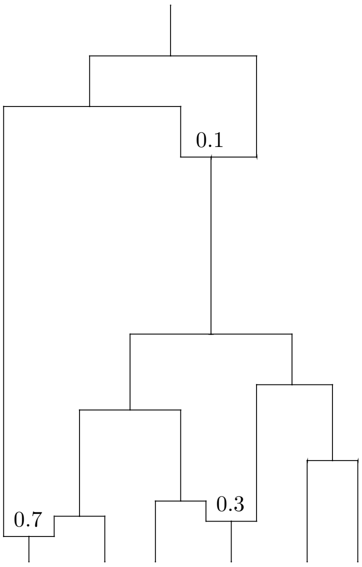

Figure 11.8. Multiple locus ancestral recombination graph.

12
Random Drift in Population Dynamics

13
Phylogenetic Trees

14 Statistics and Probability

This chapter describes subjects from statistics and probability theory that are not always included in introductory courses, but are relevant to genetics.

14.1 Contingency Tables

A *contingency table* is a vector, matrix or array of counts of individuals belonging to certain categories. A contingency table based on a random sample from a population possesses a multinomial distribution. Such tables arise frequently in genetics, and it is often desired to test whether the corresponding probability vector takes a special form. Chisquare tests, which derive their name from the asymptotic approximation to the distribution of the test statistic, are popular for this purpose. In this section we discuss the asymptotics of such tests, also giving attention to situations in which the test statistics are not asymptotically chisquared distributed. For omitted proofs we refer to Chapter 17 in Van der Vaart (1998).

14.1.1 Quadratic Forms in Normal Vectors

The *chisquare distribution* with k degrees of freedom is (by definition) the distribution of $\sum_{i=1}^k Z_i^2$ for i.i.d. $N(0,1)$-distributed variables $Z_1,\ldots,Z_k$. The sum of squares is the squared norm $\|Z\|^2$ of the standard normal vector $Z=(Z_1,\ldots,Z_k)$. The following lemma gives a characterization of the distribution of the norm of a general zero-mean normal vector.

14.1 Lemma. *If the vector X is $N_k(0,\Sigma)$-distributed, then $\|X\|^2$ is distributed as $\sum_{i=1}^k \lambda_i Z_i^2$ for i.i.d. $N(0,1)$-distributed variables $Z_1,\ldots,Z_k$ and $\lambda_1,\ldots,\lambda_k$ the eigenvalues of Σ.*

Proof. There exists an orthogonal matrix O such that $O\Sigma O^T = \operatorname{diag}(\lambda_i)$. Then the vector OX is $N_k\bigl(0, \operatorname{diag}(\lambda_i)\bigr)$-distributed, which is the same as the distribution of the vector $(\sqrt{\lambda_1}Z_1, \ldots, \sqrt{\lambda_k}Z_k)$. Now $\|X\|^2 = \|OX\|^2$ has the same distribution as $\sum(\sqrt{\lambda_i}Z_i)^2$. ■

The distribution of a quadratic form of the type $\sum_{i=1}^k \lambda_i Z_i^2$ is complicated in general. However, in the case that every λ_i is either 0 or 1, it reduces to a chisquare distribution. If this is not naturally the case in an application, then a statistic is often transformed to achieve this desirable situation. The definition of the Pearson statistic illustrates this.

14.1.2 Pearson Statistic

Suppose that we observe a vector $X_n = (X_{n,1}, \ldots, X_{n,k})$ with the multinomial distribution corresponding to n trials and k classes having probabilities $p = (p_1, \ldots, p_k)$. The *Pearson statistic* for testing the null hypothesis $H_0: p = a$ is given by

$$C_n(a) = \sum_{i=1}^k \frac{(X_{n,i} - na_i)^2}{na_i}.$$

We shall show that the sequence $C_n(a)$ converges in distribution to a chisquare distribution if the null hypothesis is true. The practical relevance is, that we can use the chisquare table to find critical values for the test. The proof shows why Pearson divided the squares by na_i, and did not propose the simpler statistic $\|X_n - na\|^2$.

14.2 Theorem. *If the vectors X_n are multinomially distributed with parameters n and $a = (a_1, \ldots, a_k) > 0$, then the sequence $C_n(a)$ converges under a in distribution to the χ^2_{k-1}-distribution.*

Proof. The vector X_n can be thought of as the sum of n independent multinomial vectors $Y_1, \ldots, Y_n$ with parameters 1 and $a = (a_1, \ldots, a_k)$. Then

$$\mathrm{E}Y_i = a, \qquad \operatorname{Cov} Y_i = \begin{pmatrix} a_1(1-a_1) & -a_1a_2 & \cdots & -a_1a_k \\ -a_2a_1 & a_2(1-a_2) & \cdots & -a_2a_k \\ \vdots & \vdots & & \vdots \\ -a_ka_1 & -a_ka_2 & \cdots & a_k(1-a_k) \end{pmatrix}.$$

By the multivariate central limit theorem, the sequence $n^{-1/2}(X_n - na)$ converges in distribution to the $N_k(0, \operatorname{Cov} Y_1)$-distribution. Consequently, with $\sqrt{a}$ the vector with coordinates $\sqrt{a_i}$,

$$\Bigl(\frac{X_{n,1} - na_1}{\sqrt{na_1}}, \ldots, \frac{X_{n,k} - na_k}{\sqrt{na_k}}\Bigr) \rightsquigarrow N(0, I - \sqrt{a}\sqrt{a}^T).$$

Since $\sum a_i = 1$, the matrix $I - \sqrt{a}\sqrt{a}^T$ has eigenvalue 0, of multiplicity 1 (with eigenspace spanned by $\sqrt{a}$), and eigenvalue 1, of multiplicity $(k-1)$ (with eigenspace equal to the orthocomplement of $\sqrt{a}$). An application of the continuous-mapping theorem and next Lemma 14.1 conclude the proof. ■

The number of degrees of freedom in the chisquared approximation for Pearson's statistic is the number of cells of the multinomial vector that have positive probability. However, the quality of the approximation also depends on the size of the cell probabilities a_j. For instance, if 1001 cells had null probabilities $10^{-23},\ldots,10^{-23},1-10^{-20}$, then it is clear that for moderate values of n all except one cells will be empty, and a huge value of n is necessary to make a χ^2_{1000}-approximation work. As a rule of thumb, it is often advised to choose the partitioning sets such that each number na_j is at least 5. This criterion depends on the (possibly unknown) null distribution, and is not the same as saying that the number of observations in each cell must satisfy an absolute lower bound, which could be very unlikely if the null hypothesis is false. The rule of thumb means to protect the level.

14.1.3 Estimated Parameters

Chisquared tests are used quite often, but usually to test more complicated hypotheses. If the null hypothesis of interest is composite, then the parameter a is unknown and cannot be used in the definition of a test statistic. A natural extension is to replace the parameter by an estimate $\hat{a}_n$ and use the statistic

$$C_n(\hat{a}_n)=\sum_{i=1}^k\frac{(X_{n,i}-n\hat{a}_{n,i})^2}{n\hat{a}_{n,i}}.$$

The estimator $\hat{a}_n$ is constructed to be a good estimator when the null hypothesis is true. The asymptotic distribution of this modified Pearson statistic is not necessarily chisquare, but depends on the estimators $\hat{a}_n$ being used. Most often the estimators will be asymptotically normal, and the statistics

$$\frac{X_{n,i}-n\hat{a}_{n,i}}{\sqrt{n\hat{a}_{n,i}}}=\frac{X_{n,i}-na_{n,i}}{\sqrt{n\hat{a}_{n,i}}}-\frac{\sqrt{n}(\hat{a}_{n,i}-a_{n,i})}{\sqrt{\hat{a}_{n,i}}}$$

will be asymptotically normal as well. Then the modified chisquare statistic will be asymptotically distributed as a quadratic form in a multivariate-normal vector. In general, the eigenvalues determining this form are not restricted to 0 or 1, and their values may depend on the unknown parameter. Then the critical value cannot be taken from a table of the chisquare distribution. There are two popular possibilities to avoid this problem.

First, the Pearson statistic is a certain quadratic form in the observations that is motivated by the asymptotic covariance matrix of a multinomial vector. When the parameter a is estimated, the asymptotic covariance matrix changes in form, and it would be natural to change the quadratic form in such a way that the resulting statistic is again chisquare distributed. This idea leads to the Rao-Robson-Nikulin modification of the Pearson statistic.

Second, we could retain the form of the Pearson statistic, but use special estimators $\hat{a}$. In particular, the maximum likelihood estimator based on the multinomial vector X_n, or the *minimum-chisquare estimator* $\bar{a}_n$ defined by, with $\mathcal{P}_0$ being the null hypothesis,

$$\sum_{i=1}^{k} \frac{(X_{n,i} - n\bar{a}_{n,i})^2}{n\bar{a}_{n,i}} = \inf_{p \in \mathcal{P}_0} \sum_{i=1}^{k} \frac{(X_{n,i} - np_i)^2}{np_i}.$$

The right side of this display is the "minimum-chisquare distance" of the observed frequencies to the null hypothesis, and is an intuitively reasonable test statistic. The null hypothesis is rejected if the distance of the observed frequency vector X_n/n to the set $\mathcal{P}_0$ is large. A disadvantage is greater computational complexity.

These two modifications, using the minimum-chisquare estimator or the maximum likelihood estimator based on X_n, may seem natural, but are artificial in some applications. For instance, in goodness-of-fit testing, the multinomial vector is formed by grouping the "raw data", and it would be more natural to base the estimators on the raw data, rather than on the grouped data. On the other hand, using the maximum likelihood- or minimum-chisquare estimator based on X_n has the advantage of a remarkably simple limit theory: if the null hypothesis is "locally linear", then the modified Pearson statistic is again asymptotically chisquare distributed, but with the number of degrees of freedom reduced by the (local) dimension of the estimated parameter.

This interesting asymptotic result is most easily explained in terms of the minimum-chisquare statistic, as the loss of degrees of freedom corresponds to a projection (i.e. a minimum distance) of the limiting normal vector. We shall first show that the two types of modifications are asymptotically equivalent, and are asymptotically equivalent to the likelihood ratio statistic as well. The likelihood ratio statistic for testing the null hypothesis $H_0: p \in \mathcal{P}_0$ is given by

$$L_n(\hat{a}_n) = \inf_{p \in \mathcal{P}_0} L_n(p), \qquad L_n(p) = 2\sum_{i=1}^{k} X_{n,i} \log \frac{X_{n,i}}{np_i}.$$

14.3 Lemma. *Let $\mathcal{P}_0$ be a closed subset of the unit simplex, and let $\hat{a}_n$ be the maximum likelihood estimator of a under the null hypothesis $H_0: a \in \mathcal{P}_0$ (based on X_n). Then*

$$\inf_{p \in \mathcal{P}_0} \sum_{i=1}^{k} \frac{(X_{n,i} - np_i)^2}{np_i} = C_n(\hat{a}_n) + o_P(1) = L_n(\hat{a}_n) + o_P(1).$$

Proof. Let $\bar{a}_n$ be the minimum-chisquare estimator of a under the null hypothesis. Both sequences of estimators $\bar{a}_n$ and $\hat{a}_n$ are $\sqrt{n}$-consistent. For the maximum likelihood estimator this follows from Corollary 5.53 in Van der Vaart (1998). The minimum-chisquare estimator satisfies by its definition

$$\sum_{i=1}^{k} \frac{(X_{n,i} - n\bar{a}_{n,i})^2}{n\bar{a}_{n,i}} \le \sum_{i=1}^{k} \frac{(X_{n,i} - na_i)^2}{na_i} = O_P(1).$$

This implies that each term in the sum on the left is $O_P(1)$, whence $n|\bar a_{n,i}-a_i|^2 = O_P(\bar a_{n,i}) + O_P\bigl(|X_{n,i}-na_i|^2/n\bigr)$ and hence the $\sqrt n$-consistency.

Next, the two-term Taylor expansion $\log(1+x)=x-\frac12 x^2+o(x^2)$ yields, for any $\sqrt n$-consistent estimator sequence $\hat p_n$,

$$\begin{aligned}\sum_{i=1}^k X_{n,i}\log\frac{X_{n,i}}{n\hat p_{n,i}} &= -\sum_{i=1}^k X_{n,i}\Bigl(\frac{n\hat p_{n,i}}{X_{n,i}}-1\Bigr)+\tfrac12\sum_{i=1}^k X_{n,i}\Bigl(\frac{n\hat p_{n,i}}{X_{n,i}}-1\Bigr)^2+o_P(1)\\ &= 0+\tfrac12\sum_{i=1}^k\frac{(X_{n,i}-n\hat p_{n,i})^2}{X_{n,i}}+o_P(1).\end{aligned}$$

In the last expression we can also replace $X_{n,i}$ in the denominator by $n\hat p_{n,i}$, so that we find the relation $L_n(\hat p_n)=C_n(\hat p_n)$ between the likelihood ratio and the Pearson statistic, for every $\sqrt n$-consistent estimator sequence $\hat p_n$. By the definitions of $\bar a_n$ and $\hat a_n$, we conclude that, up to $o_P(1)$-terms, $C_n(\bar a_n)\le C_n(\hat a_n)=L_n(\hat a_n)\le L_n(\bar a_n)=C_n(\bar a_n)$. The lemma follows. ∎

Since the minimum-chisquare estimator $\bar a_n$ (relative to $\bar{\mathcal P}_0$) is $\sqrt n$-consistent, the asymptotic distribution of the minimum-chisquare statistic is not changed if we replace $n\bar a_{n,i}$ in its denominator by the true value na_i. Next, we can decompose,

$$\frac{X_{n,i}-np_i}{\sqrt{na_i}}=\frac{X_{n,i}-na_i}{\sqrt{na_i}}-\frac{\sqrt n(p_i-a_i)}{\sqrt{a_i}}.$$

The first vector on the right converges in distribution to multivariate normal vector X as in the proof of Theorem 14.2. The (modified) minimum-chisquare statistics are the distances of these vectors to the sets $H_{n,0}=\sqrt n(\mathcal P_0-a)/\sqrt a$. If these converge in a suitably way to a limit, then the statistics ought to converge to the minimum distance of X to this set. This heuristic argument is made precise in the following theorem, which determines the limit distribution under the assumption that X_n is multinomial with parameters n and $a+g/\sqrt n$.

Say that a sequence of sets H_n *converges* to a set H if H is the set of all limits $\lim h_n$ of converging sequences h_n with $h_n\in H_n$ for every n and, moreover, the limit $h=\lim_i h_{n_i}$ of every converging subsequence h_{n_i} with $h_{n_i}\in H_{n_i}$ for every i is contained in H.

14.4 Theorem. *Let $\mathcal P_0$ be a subset of the unit simplex such that the sequence of sets $\sqrt n(\mathcal P_0-a)$ converges to a set H_0 (in $\mathbb R^k$), and suppose that $a>0$. Then, under $a+g/\sqrt n$,*

$$\inf_{p\in\mathcal P_0}\sum_{i=1}^k\frac{(X_{n,i}-np_i)^2}{np_i}\rightsquigarrow\Bigl\|X+\frac{g}{\sqrt a}-\frac{1}{\sqrt a}H_0\Bigr\|^2,$$

for a vector X with the $N(0,I-\sqrt a\sqrt a^T)$-distribution. Here $(1/\sqrt a)H_0$ is the set of vectors $(h_1/\sqrt{a_1},\ldots,h_k/\sqrt{a_k})$ as h ranges over H_0.

The value $g = 0$ in this theorem corresponds to the null hypothesis, whereas $g \neq 0$ could refer to the power at alternatives $a + g/\sqrt{n}$ that are close to the null hypothesis.

A chisquare limit distribution arises in the case that the limit set H_0 is a linear space. Under the null hypothesis this is an ordinary chisquare distribution, whereas under alternatives the chisquare distribution is noncentral. Recall that a random variable $\sum_{i=1}^k (Z_i + \delta_i)^2$ for $Z_1, \ldots, Z_k$ independent standard normal variables and $\delta = (\delta_1, \ldots, \delta_k) \in \mathbb{R}^k$ an arbitrary vector is said to possess a *noncentral chisquare distribution* with k degrees of freedom and noncentrality parameter $\|\delta\|$.

14.5 Corollary. *Let $\mathcal{P}_0$ be a subset of the unit simplex such that the sequence of sets $\sqrt{n}(\mathcal{P}_0 - a)$ converges to a linear subspace of dimension l (of $\mathbb{R}^k$), and let $a > 0$. Then both the sequence of minimum-chisquare statistics and the sequence of modified Pearson statistics $C_n(\hat{a}_n)$ converge in distribution under $a + g/\sqrt{n}$ to the noncentral chisquare distribution with $k - 1 - l$ degrees of freedom and noncentrality parameter $\big\|(I - \Pi)(g/\sqrt{a})\big\|$, for Π the orthogonal projection onto the space $(1/\sqrt{a})H_0$.*

Proof. The vector X in the preceding theorem is distributed as $Z - \Pi_{\sqrt{a}}Z$ for $\Pi_{\sqrt{a}}$ the projection onto the linear space spanned by the vector $\sqrt{a}$ and Z a k-dimensional standard normal vector. Since every element of H_0 is the limit of a multiple of differences of probability vectors, $1^T h = 0$ for every $h \in H_0$. Therefore, the space $(1/\sqrt{a})H_0$ is orthogonal to the vector $\sqrt{a}$, and $\Pi\,\Pi_{\sqrt{a}} = 0$ for Π the projection onto the space $(1/\sqrt{a})H_0$.

The distance of X to the space $(1/\sqrt{a})H_0$ is equal to the norm of $X - \Pi X$, which is distributed as the norm of $Z - \Pi_{\sqrt{a}}Z - \Pi Z$. The latter projection is multivariate-normally distributed with mean zero and covariance matrix the projection matrix $I - \Pi_{\sqrt{a}} - \Pi$ with $k - l - 1$ eigenvalues 1. The corollary for $g = 0$ therefore follows from Lemma 14.1 or 14.18.

If $g \neq 0$, then we need to take into account an extra shift $g/\sqrt{a}$. As $\langle g/\sqrt{a}, \sqrt{a}\rangle = \sum_{i=1}^k g_i = 0$, it follows that $(I - \Pi_{\sqrt{a}})(g/\sqrt{a}) = g/\sqrt{a}$. Hence the limit variable can be written as the square of $\big\|(I - \Pi)(X + g/\sqrt{a})\big\|$, which is distributed as $\big\|(I - \Pi)(I - \Pi_{\sqrt{a}})(Z + g/\sqrt{a})\big\|$. Finally we apply the result of Exercise 14.6 with $P = \Pi + \Pi_{\sqrt{a}}$. ■

14.6 EXERCISE. If Z is a standard normal vector in $\mathbb{R}^k$ and $P: \mathbb{R}^k \to \mathbb{R}^k$ an orthogonal projection onto an l-dimensional linear subspace, then $\big\|(I - P)(Z + \mu)\big\|^2$ possesses a noncentral chisquare distribution with $k - l$ degrees of freedom and noncentrality parameter $\|(I - P)\mu\|$. [Rewrite the statistic as $\sum_{i=l+1}^k \langle Z + \mu, e_i\rangle^2$ for $e_1, \ldots, e_k$ an orthonormal basis whose first l elements span the range of P.]

14.7 Example (Parametric model). If the null hypothesis is a parametric family $\mathcal{P}_0 = \{p_\theta: \theta \in \Theta\}$ indexed by a subset Θ of $\mathbb{R}^l$ with $l \leq k$ and the maps $\theta \mapsto p_\theta$ from Θ into the unit simplex are continuously differentiable homeomorphisms of full rank, then $\sqrt{n}(\mathcal{P}_0 - p_\theta) \to \dot{p}_\theta(\mathbb{R}^l)$ for every $\theta \in \overset{\circ}{\Theta}$, where $\dot{p}_\theta$ is the derivative.

Because the limit set is a linear space of dimension l, the chisquare statistics $C_n(\hat{p}_\theta)$ are asymptotically chisquare distributed with $k-l-1$ degrees of freedom.

This conclusion is immediate from Theorem 14.4 and its corollary, provided it can be shown that the sets $\sqrt{(}\mathcal{P}_0 - p_\theta)$ converge to the set $\dot{p}_\theta(\mathbb{R}^l)$ as claimed. Now the points $\theta + h/\sqrt{n}$ are contained in Θ for every $h \in \mathbb{R}^l$ and sufficiently large n and $\sqrt{n}(p_{\theta+h/\sqrt{n}} - p_\theta) \to \dot{p}_\theta h$ by the assumed differentiability of the map $\theta \mapsto p_\theta$. Furthermore, if a subsequence of $\sqrt{n}(p_{\theta_n} - p_\theta)$ converges to a point h for a given sequence $\theta_n \in \Theta$, then $\sqrt{n}(\theta_n - \theta)$ converges to $\eta = q'_{p_\theta} h$ for q the inverse map of $\theta \mapsto p_\theta$; hence $\sqrt{n}(p_{\theta_n} - p_\theta) \to \dot{p}_\theta \eta$. It follows that the sets $\sqrt{n}(\mathcal{P}_0 - p_\theta)$ converge to the range of the derivative $\dot{p}_\theta$. □

14.1.4 Nested Hypotheses

Rather than testing a null hypothesis $\mathcal{P}_0$ within the full unit simplex, one might be interested in testing it within a proper submodel of the unit simplex. More generally, we may consider a nested sequence of subsets $\mathcal{P}_0 \subset \mathcal{P}_1 \subset \cdots \subset \mathcal{P}_J$ of the unit simplex and test $\mathcal{P}_j$ as the null model within $\mathcal{P}_{j+1}$. A natural test statistic is the difference

$$C_n(\hat{a}_{n,j}) - C_n(\hat{a}_{n,j+1}),$$

for $\hat{a}_{n,j}$ the maximum likelihood estimator of the vector of success probabilities under the assumption that the true parameter belongs to $\mathcal{P}_j$. If the models $\mathcal{P}_j$ are locally linear, then these test statistics are asymptotically distributed as independent chisquare variables, and can be viewed as giving a decomposition of the discrepancy $C_n(\hat{a}_{n,0})$ between unit simplex and smallest model into discrepancies between the models. This is similar to te

14.8 Theorem. *If $a > 0$ is contained in $\mathcal{P}_0$ and the sequences of sets $\sqrt{n}(\mathcal{P}_j - a)$ converge to linear subspaces $H_j \subset \mathbb{R}^k$ of dimensions k_j, then the sequence of vectors $\bigl(C_n(\hat{a}_{n,0}) - C_n(\hat{a}_{n,1}), \ldots, C_n(\hat{a}_{n,J-1}) - C_n(\hat{a}_{n,J}), C_n(\hat{a}_{n,J}),\bigr)$ converges in distribution to a vector of independent chisquare variables, the jth variable having $k_{j+1} - k_j$ degrees of freedom (where $k_{J+1} = k$).*

Proof. By extension of Theorem 14.4 it can be shown that the sequence of statistics $\bigl(C_n(\hat{a}_{n,0}), \ldots, C_n(\hat{a}_{n,J})\bigr)$ tends in distribution to the stochastic vector $\bigl(\|X - H_0/\sqrt{a}\|^2, \ldots, \|X - H_J/\sqrt{a}\|^2\bigr)$, for $X = (I - \Pi_{\sqrt{a}})Z$ and Z a standard normal vector. As in the proof of Corollary 14.5 we have $X - \Pi_{H_j} X = (I - \Pi_j)Z$, for Π_j the orthogonal projection onto the subspace lin $\sqrt{a} + H_j/\sqrt{a}$. The result follows by representing Z as a vector of standard normal variables relative to an orthonormal basis constructed from successive orthonormal bases of the nested subspaces lin $\sqrt{a} + H_0/\sqrt{a} \subset \cdots \subset$ lin $\sqrt{a} + H_J/\sqrt{a}$. ∎

N_{11}	N_{12}	$\cdots$	N_{1r}	$N_{1.}$
N_{21}	N_{22}	$\cdots$	N_{1r}	$N_{2.}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
N_{k1}	N_{k2}	$\cdots$	N_{1r}	$N_{k.}$
$N_{.1}$	$N_{.2}$	$\cdots$	$N_{.r}$	N

Table 14.1. Classification of a population of N elements according to two categories, N_{ij} elements having value i on the first category and value j on the second. The borders give the sums over each row and column, respectively.

14.1.5 Testing Independence

Suppose that each element of a population can be classified according to two characteristics, having k and r levels, respectively. The full information concerning the classification can be given by a $(k \times r)$-*table* of the form given in Table 14.1.

Often the full information is not available, but we do know the classification $X_{n,ij}$ for a random sample of size n from the population. The matrix $X_{n,ij}$, which can also be written in the form of a $(k \times r)$-table, is multinomially distributed with parameters n and probabilities $p_{ij} = N_{ij}/N$. The null *hypothesis of independence* asserts that the two categories are independent, i.e. $H_0\colon p_{ij} = a_i b_j$ for (unknown) probability vectors a_i and b_j.

The maximum likelihood estimators for the parameters a and b (under the null hypothesis) are $\hat{a}_i = X_{n,i.}/n$ and $\hat{b}_j = X_{n,.j}/n$. With these estimators the modified Pearson statistic takes the form

$$C_n(\hat{a}_n \otimes \hat{b}_n) = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(X_{n,ij} - n\hat{a}_i\hat{b}_j)^2}{n\hat{a}_i\hat{b}_j}.$$

The null hypothesis is a $k+r-2$-dimensional submanifold of the unit simplex in $\mathbb{R}^{kr}$. In a shrinking neighbourhood of a parameter in its interior this manifold looks like its tangent space, a linear space of dimension $k+r-2$. Thus, the sequence $C_n(\hat{a}_n \otimes \hat{b}_n)$ is asymptotically chisquare distributed with $kr - 1 - (k + r - 2) = (k-1)(r-1)$ degrees of freedom.

14.9 Corollary. *If the $(k \times r)$ matrices X_n are multinomially distributed with parameters n and $p_{ij} = a_i b_j > 0$, then the sequence $C_n(\hat{a}_n \otimes \hat{b}_n)$ converges in distribution to the $\chi^2_{(k-1)(r-1)}$-distribution.*

Proof. The map $(a_1, \ldots, a_{k-1}, b_1, \ldots, b_{r-1}) \to (a \times b)$ from $\mathbb{R}^{k+r-2}$ into $\mathbb{R}^{kr}$ is continuously diffferentiable and of full rank. The true values $(a_1, \ldots, a_{k-1}, b_1 \ldots, b_{r-1})$ are interior to the domain of this map. Thus the sequence of sets $\sqrt{n}(\mathcal{P}_0 - a \times b)$ converges to a $(k + r - 2)$-dimensional linear subspace of $\mathbb{R}^{kr}$. ■

14.1.6 Comparing Two Tables

Suppose we wish to compare two contingency tables giving the classifications of two populations, based on two independent random samples from the populations. Given independent vectors X_m and Y_n with multinomial distributions with parameters m and $p = (p_1, \ldots, p_k)$ and n and $q = (q_1, \ldots, q_k)$, respectively, we wish to test the null hypothesis $H_0 \colon p = q$ that the relative frequencies in the populations are the same.

This situation is almost the same as testing independence in a $(2 \times k)$-table (with rows X_m and Y_n), the difference being that the counts of the two rows of this table are fixed to m and n in advance, and not binomial variables as in the $(2 \times k)$-table.

It is natural to base a test on the difference X_m/m and Y_n/n of the maximum likelihood estimators of p and q. Under the null hypothesis the maximum likelihood estimator of the common parameter $p = q$ is $(X_m + Y_n)/(m + n)$. A natural test statistic is therefore

$$\sum_{i=1}^{k} \frac{mn(X_{m,i}/m - Y_{n,i}/n)^2}{X_{m,i} + Y_{n,i}}. \tag{14.10}$$

The norming of the terms by the constant mn may look odd at first, but has been chosen to make the statistic asymptotically chisquare.

Another way to approach the problem would be to consider first the test statistic we would use if the value of $p = q$ under the null hypothesis were known. If this value is denoted a, then a natural test statistic is

$$C_{m,n}(a) = \sum_{i=1}^{k} \frac{(X_{m,i} - ma_i)^2}{ma_i} + \sum_{i=1}^{k} \frac{(Y_{n,i} - na_i)^2}{na_i}.$$

This is the sum of two chisquare statistics for testing $H_0 \colon p = a$ and $H_0 \colon q = a$ using X_n and Y_n, respectively. Because the common value a is not known we replace it by its maximum likelihood estimator under the null hypothesis, which is $\hat{a} = (X_m + Y_n)/(m + n)$. From Theorem 14.2 it follows that the test statistic with the true value of a is asymptotically chisquare with $2(k-1)$ degrees of freedom, under the null hypothesis. The test statistic $C_{m,n}(\hat{a})$ with the estimated cell frequencies $\hat{a}$ turns out to be asymptotically chisquare with $k - 1$ degrees of freedom, under the null hypothesis, i.e. $k - 1$ degrees of freedom are "lost".

The statistic (14.10) turns out to be algebraically identical to $C_{m,n}(\hat{a})$.

14.11 Theorem. *If the vectors X_m and Y_n are independent and multinomially distributed with parameters m and $a + g/\sqrt{m+n}$ and n and $a + h/\sqrt{m+n}$ for $a = (a_1, \ldots, a_k) > 0$ and arbitrary (g, h), then the statistic $C_{m,n}(\hat{a})$ converges as $m, n \to \infty$ such that $m/(m+n) \to \lambda \in (0,1)$ in distribution to the noncentral χ^2_{k-1}-distribution with noncentrality parameter $\|(g-h)/\sqrt{a}\|\sqrt{\lambda(1-\lambda)}$.*

Proof. First consider the case that $g=h=0$. The statistic $C_{m,n}(\hat{a})$ can be written in the form (14.10). We can decompose

$$\sqrt{m+n}\frac{X_m/m-Y_n/n}{\sqrt{a}}=\sqrt{\frac{m+n}{m}}\,\frac{(X_m-ma)}{\sqrt{ma}}-\sqrt{\frac{m+n}{n}}\,\frac{(Y_n-na)}{\sqrt{na}}.$$

As shown in the proof of Theorem 14.2 the two random vectors on the right side converge in distribution to the $N_k(0,I-\sqrt{a}\sqrt{a}^T)$-distribution. If $m/(m+n)\to\lambda$ and $n/(m+n)\to 1-\lambda$, then in view of the independence of the two vectors, the whole expression tends in distribution to a $N_k\bigl(0,(\lambda^{-1}+(1-\lambda)^{-1})(I-\sqrt{a}\sqrt{a}^T\bigr)$-distribution. Again as in the proof of Theorem 14.2, the square norm of the preceding display tends to $(\lambda(1-\lambda))^{-1}$ times a variable with the chisquare distribution with $k-1$ degrees of freedom. This square norm is the statistic $C_{m,n}(\hat{a})$ up to the scaling factor $mn/(m+n)^2$, which compensates for the multiplication by $(\lambda(1-\lambda))^{-1/2}$, and replacing a in the denominator by $(X_m+Y_n)/(m+n)$.

If g or h is not zero, then the same arguments show that the left side of the preceding display converges in distribution to

$$\frac{1}{\sqrt{\lambda(1-\lambda)}}(I-\Pi_{\sqrt{a}})Z+\frac{g-h}{\sqrt{a}},$$

for a standard normal vector Z and $\Pi_{\sqrt{a}}$ the orthogonal projection onto the linear space spanned by the vector $\sqrt{a}$. Now $\Pi_{\sqrt{a}}\bigl((h-g)/\sqrt{a}\bigr)=0$, because the coordinates of both g and h add up to 0. Finally we apply the result of Exercise 14.6. ∎

The theorem with $g=h=0$ gives the asymptotic null distribution of the test statistics, which is ordinary (central) chisquare with $k-1$ degrees of freedom. More generally, the theorem shows that the power at alternatives $(a+g/\sqrt{m+n},a+h/\sqrt{m+n})$ is determined by a noncentral chisquare distribution with an equal number of degrees of freedom but with noncentrality parameter proportional to $\bigl\|(g-h)/\sqrt{a}\bigr\|$. These alternative tend to the point (a,a) in the null hypothesis, and therefore this parameter refers to a *local power* of the test. However, the result may be loosely remembered as that the power at alternatives (a,b) is determined by square noncentrality parameter

$$\frac{mn}{m+n}\Bigl\|\frac{a-b}{\sqrt{(a+b)/2}}\Bigr\|^2.$$

14.12 Example (Comparing two binomials). For $k=2$ the test statistic compares the success probabilities p_1 and q_1 of two independent binomial variables $X_{m,1}$ and $Y_{n,1}$, with parameters (m,p_1) and (n,q_1), respectively. Because $X_{m,2}/m-Y_{n,2}/n=-(X_{m,1}/m-Y_{n,1}/n)$, the numerators of the two terms in the sum (14.10) are identical. Elementary algebra allows to rewrite the test statistic as

$$\frac{mn}{m+n}\frac{(X_{m,1}/m-Y_{n,1}/n)^2}{\hat{a}_1(1-\hat{a}_1)},$$

for $\hat{a}_1 = (X_{m,1} + Y_{n,1})/(m+n)$ the maximum likelihood estimator of the common success probability under the null hypothesis.

It is easy to verify directly, from the Central limit theorem, that this statistic is asymptotically chisquare distributed with one degree of freedom, under the null hypothesis $H_0: p_1 = q_1$. Furthermore, for alternatives with $p_1 - q_1 = h/\sqrt{m+n}$ the sequence of test statistics is asymptotically noncentrally chisquared distributed with one degree of freedom and square noncentrality parameter $\lambda(1-\lambda)h^2/\bigl(a_1(1-a_1)\bigr)$. In terms of the original parameters the square noncentrality parameter is $mn/(m+n)\,(p_1-q_1)^2/(a_1a_2)$. These assertions can of course also be obtained from the general result.

In the *case-control* setting the noncentrality parameter can be written in attractive alternative form. Suppose that a population consists of affected and nonaffected individuals, and $X_{m,1}$ and $Y_{n,1}$ are the numbers individuals that possess a certain characteristic M in independent random samples of m affected individuals and n unaffected individuals. Let $p_{M,A}, p_{M,U}, p_M, p_A, p_U$ the fractions of individuals in the population that have characteristic M and are affected, have characteristic M and are not affected, have characteristic M, etc., and let $p_{M|A} = p_{M,A}/p_A$ and $p_{M|U} = p_{M,U}/p_U$ be the corresponding conditional probabilities. The null hypothesis of interest is $H_0: p_{M|A} = p_{M|U}$, and the corresponding noncentrality parameter is

$$\frac{mn}{m+n}\frac{(p_{M|A} - p_{M|U})^2}{p_M(1-p_M)} = \frac{r_{M,A}^2}{p_Ap_U}.$$

In the last expression $r_{M,A}$ is the correlation between the 1_M and 1_U for M and U being the event that a randomly chosen individual has characteristic M or is affected, and the last equality follows after some algebra. □

14.13 EXERCISE. Verify the last formula in the preceding example. [Hint: $r_{M,A} = (p_{M,A} - p_Mp_A)/\sqrt{p_M(1-p_M)p_Ap_U}$. Write $p_{M|A} = r_{M,A}\sqrt{p_M(1-p_M)p_U/p_A} + p_M$ and $p_{M|A} = r_{M,U}\sqrt{p_M(1-p_M)p_A/p_U} + p_M$, and note that $r_{M,A} = -r_{M,U}$.]

14.2 Likelihood Ratio Statistic

Given observed data $X^{(n)}$ with probability density $p_\theta^{(n)}$ indexed by an unknown parameter θ ranging over a set Θ, the *likelihood ratio statistic* statistic for testing the null hypothesis $H_0: \theta \in \Theta_0$ versus the alternative $H_1: \theta \in \Theta - \Theta_0$ is defined as

$$\frac{\sup_{\theta\in\Theta} p_\theta^{(n)}(X^{(n)})}{\sup_{\theta\in\Theta_0} p_\theta^{(n)}(X^{(n)})} = \frac{p_{\hat{\theta}}^{(n)}(X^{(n)})}{p_{\hat{\theta}_0}^{(n)}(X^{(n)})},$$

for $\hat{\theta}$ and $\hat{\theta}_0$ the maximum likelihood estimators under the full model Θ and the null hypothesis Θ_0, respectively. In standard situations (local asymptotic normality, open Euclidean parameter spaces) twice the log likelihood ratio statistics are under the null hypothesis asymptotically distributed as a chisquare variable with degrees of freedom equal to the difference in dimensions of Θ and Θ_0. However, this result fails if the null parameter is on the boundary of the parameter set, a situation that is common in statistical genetics. In this section we give an heuristic derivation of a more general result, in the case of replicated data. Furthermore, we consider the likelihood ratio statistic based on missing data.

14.2.1 Asymptotics

We consider the situation that the observed data is a random sample $X_1, \ldots, X_n$ from a density p_θ, so that the likelihood of $X^{(n)} = (X_1, \ldots, X_n)$ is the product density $\prod_{i=1}^n p_\theta(X^i)$. The parameter set Θ is assumed to be a subset $\Theta \subset \mathbb{R}^k$ of k-dimensional Euclidean space. We consider the distribution of the likelihood ratio statistic under the "true" parameter ϑ, which is assumed to be contained in Θ_0. Introducing the local parameter spaces $H_n = \sqrt{n}(\Theta - \vartheta)$ and $H_{n,0} = \sqrt{n}(\Theta_0 - \vartheta)$, we can write two times the log likelihood ratio statistic in the form

$$\Lambda_n = 2 \sup_{h \in H_n} \log \prod_{i=1}^n \frac{p_{\vartheta + h/\sqrt{n}}}{p_\vartheta}(X^i) - 2 \sup_{h \in H_{n,0}} \log \prod_{i=1}^n \frac{p_{\vartheta + h/\sqrt{n}}}{p_\vartheta}(X^i).$$

Let $\dot\ell_\theta(x)$ and $\ddot\ell_\theta(x)$ be the first two (partial) derivatives of the map $\theta \mapsto \log p_\theta(x)$. A Taylor expansion suggests the approximation

$$\log \frac{p_{\vartheta + h/\sqrt{n}}}{p_\vartheta}(x) \approx \frac{1}{\sqrt{n}} h^T \dot\ell_\vartheta(x) - \tfrac{1}{2} \frac{1}{n} h^T \ddot\ell_\vartheta(x) h.$$

The error in this approximation is $o(n^{-1})$. This suggests that

$$\log \prod_{i=1}^n \frac{p_{\vartheta + h/\sqrt{n}}}{p_\vartheta}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^T \dot\ell_\vartheta(X_i) - \tfrac{1}{2} \frac{1}{n} \sum_{i=1}^n h^T \ddot\ell_\vartheta(X_i) h + \cdots, \tag{14.14}$$

where the remainder (the dots) is asymptotically negligible. By the Central Limit Theorem, the Law of Large Numbers, and the "Bartlett identities" $\mathrm{E}_\theta \dot\ell_\theta(X_1) = 0$ and $\mathrm{E}_\theta \ddot\ell_\theta(X_1) = -\operatorname{Cov}_\theta\bigl(\dot\ell_\theta(X_1)\bigr)$,

$$\begin{aligned} \Delta_{n,\vartheta} &:= \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot\ell_\vartheta(X_i) \overset{\vartheta}{\rightsquigarrow} N(0, I_\vartheta), \\ I_{n,\vartheta} &:= -\frac{1}{n} \sum_{i=1}^n \ddot\ell_\vartheta(X_i) \xrightarrow{\vartheta} I_\vartheta := \operatorname{Cov}_\vartheta\bigl(\dot\ell_\vartheta(X_1)\bigr). \end{aligned}$$

This suggests the approximations

$$\begin{aligned} \Lambda_n &\approx 2 \sup_{h \in H_n} \bigl(h^T \Delta_{n,\vartheta} - \tfrac{1}{2} h^T I_\vartheta h\bigr) - 2 \sup_{h \in H_{n,0}} \bigl(h^T \Delta_{n,\vartheta} - \tfrac{1}{2} h^T I_\vartheta h\bigr) \\ &= \bigl\| I_\vartheta^{-1/2} \Delta_{n,\vartheta} - I_\vartheta^{1/2} H_{n,0} \bigr\|^2 - \bigl\| I_\vartheta^{-1/2} \Delta_{n,\vartheta} - I_\vartheta^{1/2} H_n \bigr\|^2. \end{aligned}$$

Here $\|\cdot\|$ is the Euclidean distance, and $\|x-H\| = \inf\{\|x-h\|: h \in H\}$ is the distance of x to a set H. If the sets $H_{n,0}$ and H_n converge to limit sets H_0 and H in an appropriate sense, then this suggests that two times the log likelihood ratio statistic is asymptotically distributed as

$$\begin{aligned}\Lambda &= \bigl\|I_\vartheta^{-1/2}\Delta_\vartheta - I_\vartheta^{1/2}H_0\bigr\|^2 - \bigl\|I_\vartheta^{-1/2}\Delta_\vartheta - I_\vartheta^{1/2}H\bigr\|^2, \\ &= \bigl\|X - I_\vartheta^{1/2}H_0\bigr\|^2 - \bigl\|X - I_\vartheta^{1/2}H\bigr\|^2,\end{aligned} \tag{14.15}$$

for Δ_ϑ a random vector with a normal $N_k(0, I_\vartheta)$-distribution, and $X = I_\vartheta^{-1/2}\Delta_\vartheta$ possessing the $N_k(0, I)$ distribution. Below we study the distribution of the random variable on the right for a number of examples of hypotheses H and H_0.

The following theorem makes the preceding informal derivation rigorous under mild regularity conditions. It uses the following notion of *convergence of sets.* Write $H_n \to H$ if H is the set of all limits $\lim h_n$ of converging sequences h_n with $h_n \in H_n$ for every n and, moreover, the limit $h = \lim_i h_{n_i}$ of every converging sequence h_{n_i} with $h_{n_i} \in H_{n_i}$ for every i is contained in H.♯

14.16 Theorem. *Suppose that the map $\theta \mapsto p_\theta(x)$ is continuously differentiable in a neighbourhood of ϑ for every x with derivative $\dot\ell_\theta(x)$ such that the map $\theta \mapsto I_\theta = \mathrm{Cov}_\theta\bigl(\dot\ell_\theta(X_i)\bigr)$ is well defined and continuous, and such that I_ϑ is nonsingular. Furthermore, suppose that for every θ_1 and θ_2 in a neighbourhood of ϑ and for a measurable function $\dot\ell$ such that $\mathrm{E}_\vartheta\dot\ell^2(X_1) < \infty$,*

$$\bigl|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)\bigr| \le \dot\ell(x)\,\|\theta_1 - \theta_2\|.$$

If the maximum likelihood estimators $\hat\theta_{n,0}$ and $\hat\theta_n$ converge under ϑ in probability to ϑ and the sets $H_{n,0}$ and H_n converge to sets H_0 and H, then the sequence of likelihood ratio statistics Λ_n converges under $\vartheta + h/\sqrt{n}$ in distribution to the random variable Λ given in (14.15), for Δ_ϑ normally distributed with mean $I_\vartheta h$ and covariance matrix I_ϑ.

14.17 Example. If Θ_0 is the single point ϑ, then $H_0 = \{0\}$. The limit variable is then $\Lambda = \|X\|^2 - \bigl\|X - I_\vartheta^{1/2}H\bigr\|^2$. □

If ϑ is an inner point of Θ, then the set H is the full space $\mathbb{R}^k$ and the second term on the right of (14.15) is zero. If ϑ is also a relative inner point of Θ_0, then H_0 will be a linear subspace of $\mathbb{R}^k$. The following lemma then shows that the asymptotic null distribution of the likelihood ratio statistic is chisquare with $k - l$ degrees of freedom, for l the dimension of H_0.

♯ For a proof of the theorem see Van der Vaart (1998).

14.18 Lemma. *Let X be a k-dimensional random vector with a standard normal distribution and let H_0 be an l-dimensional linear subspace of $\mathbb{R}^k$. Then $\|X - H_0\|^2$ is chisquare distributed with $k - l$ degrees of freedom.*

Proof. Take an orthonormal base of $\mathbb{R}^k$ such that the first l elements span H_0. By Pythagoras' theorem, the squared distance of a vector z to the space H_0 equals the sum of squares $\sum_{i>l} z_i^2$ of its last $k - l$ coordinates with respect to this basis. A change of base corresponds to an orthogonal transformation of the coordinates. Since the standard normal distribution is invariant under orthogonal transformations, the coordinates of X with respect to any orthonormal base are independent standard normal variables. Thus $\|X - H_0\|^2 = \sum_{i>l} X_i^2$ is chisquare distributed. ■

If ϑ is a boundary point of Θ or Θ_0, then the limit sets H or H_0 will not be linear spaces, and the limit distribution is typically not chisquare.

14.19 Example (Recombination fraction). A recombination fraction θ between two loci is known to belong to the interval $\Theta = [0, \frac{1}{2}]$. To test whether a disease locus is linked to a marker locus we want to the test the null hypothesis $H_0: \theta = \frac{1}{2}$, which is a boundary point of the parameter set. The set $H_n = \sqrt{n}(\Theta - \frac{1}{2})$ is equal to $[-\frac{1}{2}\sqrt{n}, 0]$ and can be seen to converge to the half line $H = (-\infty, 0]$. Under the assumption that the Fisher information is positive, the set $I_{1/2}^{1/2} H$ is the same half line***.

The asymptotic null distribution of the log likelihood ratio statistic is the distribution of $|X|^2 - |X - H|^2$ for X a standard normal variable. If $X > 0$, then 0 is the point in H that is closest to X and hence $|X|^2 - |X - H|^2 = 0$. If $X \leq 0$, then X is itself the closest point and hence $|X|^2 - |X - H|^2 = X^2$, which is χ^2-distributed with one degree of freedom. Because the normal distribution is symmetric, these two possibilities occur with probability $\frac{1}{2}$. We conclude that the limit distribution of 2 times the log likelihood ratio statistic is a mixture of a point mass at zero and a chisquare distribution with one degree of freedom.

For $\alpha < \frac{1}{2}$ the equation $\frac{1}{2}P(\xi_1^2 \geq c) + \frac{1}{2}P(0 \geq c) = \alpha$ is equivalent to $P(\xi_1^2 \geq c) = 2\alpha$. Hence the upper α-quantile of this mixture distribution is the upper 2α-quantile of the chisquare distribution with one degree of freedom. □

14.20 Example (Half spaces). Suppose that the parameter set Θ is a halfspace $\Theta = \{(\alpha, \beta): \alpha \in \mathbb{R}^k, \beta \in \mathbb{R}, \beta \geq 0\}$ and the null hypothesis is $H_0: \beta = 0$, i.e. $\Theta_0 = \{(\alpha, \beta): \alpha \in \mathbb{R}^k, \beta = 0\}$. Under the assumption that $\vartheta \in \Theta_0$ the limiting local parameter spaces corresponding to Θ and Θ_0 are $H = \mathbb{R}^k \times [0, \infty)$ and $H_0 = \mathbb{R}^k \times \{0\}$.

The image of a half space $\{x: x^T n \leq 0\}$ under a nonsingular linear transformation A is $\{Ax: x^T n \leq 0\} = \{y: y^T (A^{-1})^T n \leq 0\}$, which is again a half space Therefore, for a strictly positive-definite matrix $I_\vartheta^{1/2}$ the image $H^1 = I_\vartheta^{1/2} H$ is again a halfspace, and its boundary hyperplane is $H_0^1 = I_\vartheta^{1/2} H_0$.

By the rotational symmetry of the standard normal distribution, the distribution of the variables $\|X - H_0^1\|^2 - \|X - H^1\|^2$ does not depend on the orientation of the halfspace H^1 and space H_0^1. Taking these spaces as H_0 and H_1, it is therefore not difficult to see that this variable has the same distribution as in Example 14.19. □

14.21 Example (Holmans' triangle). Holmans' triangle as shown in Figure 5.1 has as limiting set H a convex cone with apex at 0. The set $I_\vartheta^{1/2}H$ is also a convex cone, strictly contained in a halfspace.

Indeed, the geometric action of a positive-definite matrix is to rescale (by multiplication with the eigenvalues) the coordinates relative to the basis of eigenvectors. The four quadrants spanned by the eigenvalues are left invariant. If the triangle H is contained in a quadrant, then so is the set $I_\vartheta^{1/2}H$, whence its boundary lines make an angle of less than 90 degrees. If the triangle H covers parts of two quadrants, then its image still remains within the union of these quadrants and hence its boundary lines make an angle of less than 180 degrees.

Figure 14.1 gives an example of a limit set $I_\vartheta^{1/2}H$ in two dimensions. The variable $X = I_\vartheta^{1/2}\Delta_\vartheta$ is standard normally distributed. In this example the limit distribution is a mixture of chisquare distributions, because

$$\|X\|^2 - \|X - A\|^2 = \begin{cases} \|X\|^2, & \text{if } X \in A, \\ 0, & \text{if } X \in C, \\ \|\Pi_1 X\|^2, & \text{if } X \in B, \\ \|\Pi_2 X\|^2, & \text{if } X \in D, \end{cases},$$

where Π_1 and Π_2 are the projections onto the lines perpendicular to the boundary lines of A. □

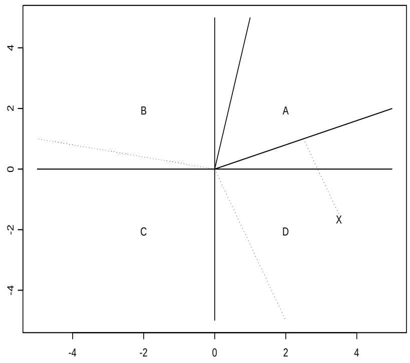

Figure 14.1. The area A refers to the set $I_\vartheta^{1/2}H$. Indicated is the projection of a point X in area D onto the set A. The projection of a vector X onto A is equal to X if $X \in A$; it is 0 if $X \in C$ and it is on the boundary of A if X is in B or D.

* 14.2.2 Asymptotics with Missing Data

Suppose that the random variable (X, Y) follows a statistical model given by the density r_θ, but we observe only X, so that the marginal density p_θ of X provides the relevant likelihood. The following lemma shows how the likelihood ratio statistic for observing X can be computed from the likelihood ratio statistic for observing (X, Y) by a conditional expectation. Assume that the distribution of (X, Y) under θ is absolutely continuous with respect to the distribution under θ_0, so that the likelihood ratio is well defined.

14.22 Lemma. *If X and Y are random variables with joint density r_θ and $r_\theta \ll r_{\theta_0}$, then the marginal density p_θ of X satisfies*

$$\frac{p_\theta(X)}{p_{\theta_0}(X)} = \mathrm{E}_{\theta_0}\Bigl(\frac{r_\theta(X,Y)}{r_{\theta_0}(X,Y)} \,|\, X\Bigr).$$

Proof. We use the equality $\mathrm{E}\mathrm{E}(Z|\,X)f(X) = \mathrm{E}Zf(X)$, which is valid for any random variables X and Z and every measurable function f. With $h_\theta(X)$ the conditional expectation in the right side of the lemma and $Z = (r_\theta/r_{\theta_0})(X, Y)$, this yields

$$\begin{aligned}\mathrm{E}_{\theta_0} h_\theta(X) f(X) &= \mathrm{E}_{\theta_0} \frac{r_\theta(X,Y)}{r_{\theta_0}(X,Y)} f(X)\\ &= \int \frac{r_\theta(x,y)}{r_{\theta_0}(x,y)} f(x)\, r_{\theta_0}(x,y)\, d\mu(x,y)\\ &= \int r_\theta(x,y) f(x)\, d\mu(x,y) = \mathrm{E}_\theta f(X).\end{aligned}$$

It follows that $h_\theta f p_{\theta_0} = f p_\theta$ almost everywhere under θ_0 for any f, which implies that $h_\theta p_{\theta_0} = p_\theta$. ∎

Next consider the situation that we would have liked to base a test on the likelihood ratio statistic for observing Y, but we only observe X. If s_θ is the density of Y under θ, then it seems intuitively reasonable to use the statistic

$$\mathrm{E}_{\theta_0}\Bigl(\frac{s_\theta(Y)}{s_{\theta_0}(Y)} \,|\, X\Bigr). \tag{14.23}$$

However, this is not necessarily the likelihood ratio statistic for observing X. The preceding lemma does not apply, since we do not condition the unobserved variable on a subset.

To save the situation we can still interpret the preceding display as a likelihood ratio statistic, by introducing an artificial model for the variable (X, Y) as follows. Given the conditional density $(x, y) \mapsto t_{\theta_0}(x|\,y)$ of X given Y under θ_0, make the working hypothesis that (X, Y) is distributed according to the density

$$(x, y) \mapsto t_{\theta_0}(x|\,y) s_\theta(y).$$

Even though there may be a reasonable model under which the law of X given Y depends on the parameter θ, we adopt this artificial model, in which the conditional law is fixed. The likelihood ratio for observing (X,Y) given the artificial model is

$$\frac{t_{\theta_0}(X|\,Y)s_\theta(Y)}{t_{\theta_0}(X|\,Y)s_{\theta_0}(Y)}=\frac{s_\theta(Y)}{s_{\theta_0}(Y)}.$$

Thus if we apply the conditioning procedure of the preceding lemma to the artificial model, we obtain exactly the statistic (14.23).

Using the "true conditional density" t_θ, and hence the true likelihood ratio test based on the observed data X, may yield a more powerful test. However, at least the statistic (14.23) can be interpreted as a likelihood ratio statistic in some model, and hence general results for likelihood ratio statistics should apply to it. In particular, the distribution of (X,Y) under θ_0 is the same in the artificial and correct model. We conclude that the statistic (14.23) therefore behaves under the null hypothesis the same as the likelihood ratio statistic based on observing X from the model in which (X,Y) is distributed according to $(x,y)\mapsto t_{\theta_0}(x|\,y)s_\theta(y)$, i.e. for X having density $x\mapsto\int t_{\theta_0}(x|\,y)s_\theta(y)\,d\nu(y)$. This reduces under θ_0 to the true marginal distribution of X and hence under the null hypothesis the statistic (14.23) has the distribution as in Theorem 14.16, where the Fisher information matrix must be computed for the model given by the densities $x\mapsto\int t_{\theta_0}(x|\,y)s_\theta(y)\,d\nu(y)$. If H is a linear space, then the null limit distribution is chisquare.

Theorem 14.16 does not yield the relevant limit distribution under alternatives. The power of the test statistic (14.23) is the same as the power of the likelihood ratio statistic based on X having density q_θ given by

$$q_\theta(x)=\int t_{\theta_0}(x|\,y)s_\theta(y)\,dy. \tag{14.24}$$

This is not the power of the likelihood ratio statistic based on X, because the model q_θ is misspecified.

The following theorem extends Theorem 14.16 to this situation. Suppose that the observation X is distributed according to a density p_θ, but we use the likelihood ratio statistic for testing $H_0{:}\,\theta\in\Theta_0$ based on the assumption that X has density q_θ.

14.25 Theorem. *Assume that $\vartheta\in\Theta_0$ with $q_\vartheta=p_\vartheta$. Suppose that the maps $\theta\mapsto p_\theta(x)$ and $\theta\mapsto q_\theta(x)$ are continuously differentiable in a neighbourhood of ϑ for every x with derivatives $\dot\ell_\theta(x)$ and $\dot\kappa_\theta(x)$ such that the maps $\theta\mapsto I_\theta=\operatorname{Cov}_\theta\bigl(\dot\ell_\theta(X_i)\bigr)$ and $\theta\mapsto J_\theta=\operatorname{Cov}_\theta\bigl(\dot\kappa_\theta(X_i)\bigr)$ are well defined and continuous, and such that J_ϑ is nonsingular. Furthermore, suppose that for every θ_1 and θ_2 in a neighbourhood of ϑ and for a measurable function $\dot\kappa$ such that $P_\vartheta\dot\kappa^2<\infty$,*

$$\bigl|\log q_{\theta_1}(x)-\log q_{\theta_2}(x)\bigr|\le\dot\kappa(x)\,\|\theta_1-\theta_2\|.$$

If the maximum likelihood estimators $\hat\theta_{n,0}$ and $\hat\theta_n$ for the model $\{q_\theta: \theta \in \Theta\}$ are consistent under ϑ and the sets $H_{n,0}$ and H_n converge to sets H_0 and H, then the sequence of likelihood ratio statistics Λ_n converges under $\vartheta + h/\sqrt{n}$ in distribution to Λ given in (14.15), for Δ_ϑ normally distributed with mean $\mathrm{Cov}_\vartheta(\dot\kappa_\vartheta, \dot\ell_\vartheta)h$ and covariance matrix J_ϑ.

In the present case the density q_θ takes the form (14.24). The score function at θ_0 for this model is

$$\dot\kappa_{\theta_0} = \mathrm{E}_{\theta_0}\Bigl(\frac{\dot s_{\theta_0}(Y)}{s_{\theta_0}(Y)} \,|\, X\Bigr).$$

The Fisher information J_{θ_0} is the covariance matrix of this. The covariance appearing in the theorem is

$$\mathrm{Cov}_\vartheta(\dot\kappa_\vartheta, \dot\ell_\vartheta) = \mathrm{E}_{\theta_0}\Bigl(\frac{\dot s_{\theta_0}(Y)}{s_{\theta_0}(Y)}\Bigr)\dot\ell_{\theta_0}(X).$$

14.3 Score Statistic

Implementation of the likelihood ratio test requires the determination of the maximum likelihood estimator both under the full model and under the null hypothesis. This can be computationally intensive. The *score test* is an alternative that requires less computation and provides approximately the same power when the number of observations is large. The score test requires the computation of the maximum likelihood estimator under the null hypothesis, but not under the full model. It is therefore particularly attractive when the same null hypothesis is tested versus multiple alternatives. Genome scans in genetics provide an example of this situation.

The score function of a statistical model given by probability densities p_θ indexed by a parameter $\theta \in \mathbb{R}^k$ is defined as the gradient $\dot\ell_\theta(x) = \nabla_\theta \log p_\theta(x)$ of the log density relative to the parameter. Under regularity conditions it satisfies $\mathrm{E}_\theta \dot\ell_\theta(X) = 0$, for every parameter θ. Therefore, a large deviation of $\dot\ell_\vartheta(X)$ from 0 gives an indication that ϑ is not the parameter value that has produced the data X. The principle of the score test is to reject the null hypothesis $H_0: \theta = \vartheta$ if the score function $\dot\ell_\vartheta(X)$ evaluated at the observation is significantly different from 0. For a composite null hypothesis $H_0: \theta \in \Theta_0$ the value ϑ is replaced by its maximum likelihood estimator $\hat\theta_0$ under the null hypothesis, and the test is based on $\dot\ell_{\hat\theta_0}(X)$.

A different intuition is to think of this statistic as arising in an approximation to the likelihood ratio statistic:

$$\log \frac{p_{\hat\theta}(X)}{p_{\hat\theta_0}(X)} \approx (\hat\theta - \hat\theta_0)^T \dot\ell_{\hat\theta_0}(X).$$

If the score statistic $\dot\ell_{\hat\theta_0}(X)$ is significantly different from zero, then this approximation suggests that the likelihood ratio between full and alternative hypothesis is large. To make this precise, it is necessary to take also the directions of the estimators and the second order term of the expansion into account. However, the suggestion that the likelihood ratio statistic and score statistic are closely related is correct, as shown below.

The problem is to quantify "significantly different from 0". We shall consider this in the case that $X = (X_1, \ldots, X_n)$ is a random sample of identically distributed observations. Then the probability density of X takes the form $(x_1, \ldots, x_n) \mapsto \prod_{i=1}^n p_\theta(x_i)$, for p_θ the density of a single observation, and the score statistic (divided by $\sqrt{n}$) takes the form

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\hat{\theta}_0}(X_i), \tag{14.26}$$

where $\dot{\ell}_\theta$ is the score function for a single observation. To measure whether the score statistic S_n is close to zero, the *score test* uses a weighted norm: the null hypothesis is rejected for large values of the statistic

$$\Bigl\| I_{\hat{\theta}_0}^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\hat{\theta}_0}(X_i) \Bigr\|^2 = S_n^T I_{\hat{\theta}_0}^{-1} S_n. \tag{14.27}$$

Here $I_\theta = \mathrm{E}_\theta \dot{\ell}_\theta(X_i) \dot{\ell}_\theta^T(X_i)$ is the Fisher information matrix. In standard situations, where the null hypothesis is "locally" a linear subspace of dimension k_0, this statistic can be shown to be asymptotically (as $n \to \infty$) chisquare distributed with $k - k_0$ degrees of freedom under the null hypothesis. The critical value of the score test is then chosen equal to the upper α_0-quantile of the this chisquare distribution. We study the asymptotics of the score statistic in more generality below.

14.28 Example (Simple null hypothesis). If the null hypothesis $H_0{:}\, \theta = \vartheta$ is simple, then the maximum likelihood estimator $\hat{\theta}_0$ is the deterministic parameter ϑ, and the score statistic is the sum of the independent, identically distributed random variables $\dot{\ell}_\vartheta(X_i)$. The asymptotics of the score test follow from the Central Limit Theorem, which shows that, under the null hypothesis,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_\vartheta(X_i) \rightsquigarrow N_k(0, I_\vartheta).$$

The score test statistic $S_n^T I_\vartheta^{-1} S_n$ is therefore asymptotically chisquare distributed with k degrees of freedom. □

14.29 Example (Partitioned parameter). Consider the situation of a partitioned parameter $\theta = (\theta_1, \theta_2)$ and a null hypothesis of the form $\Theta_0 = \{(\theta_1, \theta_2){:}\, \theta_1 \in \mathbb{R}^{k_0}, \theta_2 = 0\}$.

The score function can be partitioned as well as $\dot{\ell}_\theta = (\dot{\ell}_{\theta,1}, \dot{\ell}_{\theta,2})$, for $\dot{\ell}_{\theta,i}$ the vector of partial derivatives of the log density with respect to the coordinates of

θ_i. The maximum likelihood estimator under the null hypothesis has the form $\hat\theta_0 = (\hat\theta_{0,1}, 0)$, for $\hat\theta_{0,1}$ a solution to the likelihood equation

$$\sum_{i=1}^n \dot\ell_{\hat\theta_0,1}(X_i) = 0.$$

This is a system of equations with as many equations as the dimension of θ_1 in $\theta = (\theta_1, \theta_2)$. The vector $\sum_{i=1}^n \dot\ell_{\hat\theta_0}(X_i)$ takes the form $(0, \sum_{i=1}^n \dot\ell_{\hat\theta_0,2}(X_i))$ and the score test statistic (14.27) reduces to

$$\frac{1}{n}\Bigl(\sum_{i=1}^n \dot\ell_{\hat\theta_0,2}(X_i)\Bigr)^T \bigl(I_{\hat\theta_0}^{-1}\bigr)_{2,2}\Bigl(\sum_{i=1}^n \dot\ell_{\hat\theta_0,2}(X_i)\Bigr). \tag{14.30}$$

Here $(I_{\hat\theta_0}^{-1})_{2,2}$ is the relevant submatrix of the inverse information matrix $I_{\hat\theta_0}^{-1}$. (Note that a submatrix $(A^{-1})_{2,2}$ of an inverse A^{-1} is not the inverse of the submatrix $A_{2,2}$.) We can interpret the statistic (14.30) as a measure of success for the maximum likelihood estimator $\hat\theta_0 = (\hat\theta_{0,1}, 0)$ under the null hypothesis to reduce the score equation $\sum_{i=1}^n \dot\ell_\theta(X_i)$ of the full model to zero. Because $\sum_{i=1}^n \dot\ell_{\hat\theta}(X_i) = 0$ for the maximum likelihood estimator $\hat\theta$ for the full model, the score statistic can also be understood as a measure of discrepancy between the maximum likelihood estimators under the null hypothesis and in the full model.

The score statistic in this example can also be related to the *profile likelihood* for the parameter of the interest θ_2. This is defined as the maximum of the likelihood over the "nuisance parameter" θ_1, for fixed θ_2:

$$\mathrm{proflik}(\theta_2) = \sup_{\theta_1} \prod_{i=1}^n p_{(\theta_1,\theta_2)}(X_i).$$

Assume that for every θ_2 the supremum is attained, at the (random) value $\hat\theta_1(q_2)$, and assume that the function $\theta_2 \mapsto \hat\theta_1(\theta_2)$ is differentiable. Then the gradient of the log profile likelihood exists, and can be computed as

$$\nabla_{\theta_2} \log \mathrm{proflik}(\theta_2) = \sum_{i=1}^n \hat\theta_1'(\theta_2)\dot\ell_{(\hat\theta_1(\theta_2),\theta_2),1}(X_i) + \sum_{i=1}^n \dot\ell_{(\hat\theta_1(\theta_2),\theta_2),2}(X_i) = \sum_{i=1}^n \dot\ell_{(\hat\theta_1(\theta_2),\theta_2),2}(X_i),$$

because $\sum_{i=1}^n \dot\ell_{(\hat\theta_1(\theta_2),\theta_2),1}(X_i)$ vanishes by the definition of $\hat\theta_1$. If we next evaluate this at the value $\theta_2 = 0$ given by the null hypothesis, we find the score statistic $\sum_{i=1}^n \dot\ell_{\hat\theta_0,2}(X_i)$, utilized in (14.30). Furthermore, it can also be shown that the negative second derivative of the profile likelihood

$$-\frac{\partial^2}{\partial\theta_2^2} \log \mathrm{proflik}(\theta_2)$$

is a consistent estimator of the inverse of the norming matrix $\bigl(I_{\hat\theta_0}^{-1}\bigr)_{2,2}$ in (14.30). Thus after "profiling out" the nuisance parameter θ_2, the profile likelihood can be used as an ordinary likelihood for θ_2. □

If the local parameter spaces $H_{n,0} = \sqrt{n}(\Theta_0 - \vartheta)$ converge to a set H_0 that is not a linear space, then the maximum likelihood estimator under the null hypothesis is not asymptotically normally distributed, and as a consequence neither is the score statistic (cf. Theorem 14.16). The usual solution is to relax the restrictions imposed by the null hypothesis so that asymptotically it becomes a linear space, or equivalently to use the solution to the likelihood equations (under the null hypothesis) rather than the maximum likelihood estimator. In any case to gain insight in the asymptotic behaviour of the score statistic we can follow similar arguments as in Section 14.2.

The log likelihood function viewed as a function of the local parameter $h = \sqrt{n}(\theta - \vartheta)$ satisfies (cf. (14.14))

$$\log \prod_{i=1}^n \frac{p_{\vartheta+h/\sqrt{n}}}{p_\vartheta}(X_i) = h^T\Delta_{n,\vartheta} - \tfrac{1}{2}h^T I_\vartheta h + \cdots, \tag{14.31}$$

where $\Delta_{n,\theta} = n^{-1/2}\sum_{i=1}^n \dot\ell_\theta(X_i)$. The maximum likelihood estimator of θ under the null hypothesis is $\hat\theta_0 = \vartheta + \hat h_0/\sqrt{n}$ for $\hat h_0$ the maximizer of this process over h ranging over the local parameter space $H_{n,0} = \sqrt{n}(\Theta_0 - \vartheta)$. The score statistic (14.26) is the gradient of the log likelihood at $\hat\theta_0$, which can be expressed in the local parameter as

$$S_n = \frac{\partial}{\partial h}\log\prod_{i=1}^n \frac{p_{\vartheta+h/\sqrt{n}}}{p_\vartheta}(X_i)\Big|_{h=\hat h_0}.$$

If the remainder (the dots) in the expansion (14.31) of the log likelihood process can be neglected, this statistic behaves as

$$\frac{\partial}{\partial h}\big(h^T\Delta_{n,\vartheta} - \tfrac{1}{2}h^T I_\vartheta h\big)\Big|_{h=\hat h_0} = \Delta_{n,\vartheta} - I_\vartheta \hat h_0.$$

In view of Theorem 14.16, if the local parameter spaces $H_{n,0}$ tend in a suitable manner to a limit set H_0, then $\hat h_0$ behaves asymptotically as the maximizer of the process

$$h \mapsto h^T\Delta_\vartheta - \tfrac{1}{2}h^T I_\vartheta h,$$

where Δ_ϑ is the limit of the sequence $\Delta_{n,\vartheta}$. Under the parameter ϑ this vector possesses a normal distribution with mean 0 and covariance matrix I_ϑ. This suggests that the score statistic is asymptotically distributed as

$$\begin{aligned}\Delta_\vartheta - I_\vartheta &\operatorname*{argmax}_{h\in H_0}(h^T\Delta_\vartheta - \tfrac{1}{2}h^T I_\vartheta h)\\ &= I_\vartheta^{1/2}\Big(X - I_\vartheta^{1/2}\operatorname*{argmin}_{h\in H_0}\|X - I_\vartheta^{1/2}h\|\Big)\\ &= I_\vartheta^{1/2}\big(X - \Pi_{I_\vartheta^{1/2}H_0}X\big),\end{aligned}$$

where $X = I_\vartheta^{-1/2}\Delta_\vartheta$, and $\Pi_A x$ denotes the point in the set A that is closest to x (assuming that it exists) in the Euclidean norm. The vector X possesses a standard normal distribution. The score test statistic is the weighted norm $S_n^T I_{\hat\theta_0}^{-1} S_n$ of the score statistic, and should be asymptotically distributed as the corresponding weighted norm of the right side of the display, where the weighting matrix $I_{\hat\theta_0}$ can of course be replaced by its limit. In other words, the preceding heuristic argument suggests that, under the null hypothesis

$$S_n^T I_{\hat\theta_0}^{-1} S_n \rightsquigarrow \|X - I_\vartheta^{1/2} H_0\|^2, \tag{14.32}$$

where X possesses a standard normal distribution.

If H_0 is a linear subspace of $\mathbb{R}^k$ of dimension k_0 and the Fisher information matrix is nonsingular, then $I_\vartheta^{1/2} H_0$ is also a k_0-dimensional linear subspace. The variable on the right side of the preceding display then possesses a chisquare distribution with $k - k_0$ degrees of freedom. In less regular cases the distribution is nonstandard.

The following theorem makes the preceding rigorous.

14.33 Theorem. *Assume that the conditions of Theorem 14.16 holds and in addition that there exists a measurable function $\ddot\ell$ such that $\mathrm{E}_\vartheta \ddot\ell^2(X_1) < \infty$ and for every θ_1 and θ_2 in a neighbourhood of ϑ*

$$\bigl\|\dot\ell_{\theta_1}(x) - \dot\ell_{\theta_2}(x)\bigr\| \le \ddot\ell(x)\,\|\theta_1 - \theta_2\|.$$

Then under $\vartheta + h/\sqrt{n}$ the score statistic (14.26) satisfies (14.32) for a vector X with a $N_k(I_\vartheta^{1/2} h, I)$-distribution. In particular, the limit distribution under ϑ is chisquare with $k - k_0$ degrees of freedom if H_0 is a k_0-dimensional subspace.

Proof. By simple algebra the score statistic can be written as

$$\begin{aligned} S_n &= \sqrt{n}\mathbb{P}_n \dot\ell_{\vartheta+\hat h_0/\sqrt{n}} \\ &= \mathbb{G}_n(\dot\ell_{\vartheta+\hat h_0/\sqrt{n}} - \dot\ell_\vartheta) + \mathbb{G}_n \dot\ell_\vartheta + (\sqrt{n} P_\vartheta \dot\ell_{\vartheta+\hat h_0/\sqrt{n}} + I_\vartheta \hat h_0) - I_\vartheta \hat h_0. \end{aligned}$$

The second and fourth terms on the right are as in the discussion preceding the theorem, and tend in distribution to the limits as asserted. It suffices to show that the first and thirds terms on the right tend to zero in probability.

For the second term this follows because the conditions imply that the class of functions $\{\dot\ell_\theta : \theta \in B\}$ for a sufficiently small neighbourhood B of ϑ is Donsker, and the map $\theta \mapsto \dot\ell_\theta$ is continuous in second mean at ϑ.

Minus the third term can be rewritten as,

$$\begin{aligned} &\int \dot\ell_{\hat\theta_0}(p_{\hat\theta_0}^{1/2} + p_\vartheta^{1/2})\bigl[\sqrt{n}(p_{\hat\theta_0}^{1/2} - p_\vartheta^{1/2}) - \tfrac12 \hat h_0^T \dot\ell_\vartheta p_\vartheta^{1/2}\bigr]\,d\mu \\ &\qquad + \int \dot\ell_{\hat\theta_0}(p_{\hat\theta_0}^{1/2} - p_\vartheta^{1/2})\tfrac12 \hat h_0^T \dot\ell_\vartheta p_\vartheta^{1/2}\,d\mu + \int (\dot\ell_{\hat\theta_0} - \dot\ell_\vartheta)\hat h_0^T \dot\ell_\vartheta p_\vartheta\,d\mu. \end{aligned}$$

These three terms can all be shown to tend to zero in probability, by using the Cauchy-Schwarz inequality and the implied differentiability in quadratic mean of the model. See the proof of Theorem 25.54 in Van der Vaart (1998) for a similar argument. ∎

For the asymptotic chisquare distribution of the score test statistic it is essential that the null maximum likelihood estimator $\hat\theta_0$ is asymptotically normal. It is also clear from the heuristic discussion that the result depends on the form of the asymptotic covariance matrix of this estimator. If in the definition of the score statistic (14.26) the unknown null parameter were estimated by another asymptotically normal estimator than the null maximum likelihood estimator $\hat\theta_0$, then the limit distribution of the score test statistic could be different from chisquare, as it would not reduce to the distribution of a square projection, as in (14.32).

It was already noted that the likelihood ratio statistic and score statistic are close relatives. The following theorem shows that in the regular case, with linear local parameter spaces, they are both asymptotically equivalent to the *Wald statistic*, which measures the difference between the maximum likelihood estimators between full and null hypothesis.

14.34 Theorem. *Assume that the conditions of Theorems 14.16 and 14.33 hold with $H = \mathbb{R}^k$ and H_0 a k_0-dimensional linear subspace of $\mathbb{R}^k$. Then, under ϑ, as $n \to \infty$, the score statistic S_n and likelihood ratio statistic Λ_n satisfy*

$$S_n^T I_\vartheta^{-1} S_n - n(\hat\theta - \hat\theta_0)^T I_\vartheta(\hat\theta - \hat\theta_0) \rightsquigarrow 0,$$

$$\Lambda_n - n(\hat\theta - \hat\theta_0)^T I_\vartheta(\hat\theta - \hat\theta_0) \rightsquigarrow 0.$$

Moreover, the sequence $n(\hat\theta - \hat\theta_0)^T I_\vartheta(\hat\theta - \hat\theta_0)$ tends in distribution to a chisquare distribution with $k - k_0$ degrees of freedom.

Proof. The maximum likelihood estimator under the full model satisfies

$$\hat h = \sqrt{n}(\hat\theta - \vartheta) = I_\vartheta^{-1}\Delta_{n,\vartheta} + o_P(1).$$

See for instance Van der Vaart (1998), Theorem 5.39. The score statistic S_n was seen to be asymptotically equivalent to $\Delta_{n,\vartheta} - I_\vartheta \hat h_0 = I_\vartheta(\hat h - \hat h_0)$. The first assertion is immediate from this.

If $H = \mathbb{R}^k$, then the likelihood ratio statistic Λ_n is asymptotically equivalent to (see (14.15) and the proof of Theorem 14.16)

$$\|I_\vartheta^{-1/2}\Delta_{n,\vartheta} - I_\vartheta^{1/2} H_0\|^2 = S_n^T I_\vartheta^{-1} S_n + o_P(1),$$

by (14.32). This proves the second (displayed) assertion.

That the sequence of Wald statistics is asymptotically chisquare now follows, because this is true for the other two sequences of test statistics. ■

14.4 Multivariate Normal Distribution

The d-dimensional multivariate normal distribution (coded as $N_d(\mu,\Sigma)$) is characterized by a mean vector $\mu\in\mathbb{R}^d$ and a $(d\times d)$-covariance matrix Σ. If Σ is positive definite, then the distribution has density function

$$x\mapsto \frac{1}{(2\pi)^{d/2}\sqrt{\det\Sigma}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}.$$

In this section we review statistical methods for the multivariate normal distribution.

The log likelihood function for observing a random sample $X_1,\ldots,X_n$ from the multivariate normal $N_d(\mu,\Sigma)$-distribution is up to the additive constant $-n(d/2)\log(2\pi)$ equal to

$$\begin{aligned}(\mu,\Sigma)\mapsto & -\tfrac{1}{2}n\log\det\Sigma-\tfrac{1}{2}\sum_{i=1}^n(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\\ &=-\tfrac{1}{2}n\log\det\Sigma-\tfrac{1}{2}\operatorname{tr}\Bigl(\Sigma^{-1}\sum_{i=1}^n(X_i-\mu)(X_i-\mu)^T\Bigr),\end{aligned}$$

where $\operatorname{tr}(A)$ is the trace of the matrix A. The last equality follows by application of the identities $\operatorname{tr}(AB)=\operatorname{tr}(BA)$ and $\operatorname{tr}(A+B)=\operatorname{tr}(A)+\operatorname{tr}(B)$, valid for any matrices A and B. The maximum likelihood estimator for (μ,Σ) is the point of maximum of this expression in the parameter space. If the parameter space is the maximal parameter space, consisting of all vectors μ in $\mathbb{R}^d$ and all positive-definite matrices Σ, then the maximum likelihood estimators are the *sample mean* and *sample covariance matrix*

$$\begin{aligned}\bar X&=\frac{1}{n}\sum_{i=1}^n X_i,\\ S&=\frac{1}{n}\sum_{i=1}^n(X_i-\bar X)(X_i-\bar X)^T.\end{aligned}$$

14.35 Lemma. *The maximum likelihood estimator for (μ,Σ) based on a random sample $X_1,\ldots,X_n$ from the $N_d(\mu,\Sigma)$-distribution in the unrestricted parameter space is $(\bar X,S)$.*

Proof. For fixed Σ maximizing the likelihood with respect to μ is the same as minimizing the quadratic form $\mu\mapsto\sum_{i=1}^n(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)$. This is a strictly convex function and hence has a unique minimum. The stationary equation is given by

$$0=\frac{\partial}{\partial\mu}\sum_{i=1}^n(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)=-2\sum_{i=1}^n\Sigma^{-1}(X_i-\mu).$$

This is solved uniquely by $\mu = \bar{X}$. As this solution gives the maximum of the likelihood with respect to μ for any given Σ, the absolute maximum of the likelihood is achieved at $(\bar{X}, \hat{\Sigma})$ for some $\hat{\Sigma}$.

The maximum likelihood estimator for Σ can now be found by maximizing the likelihood with respect to Σ for μ fixed at $\bar{X}$. This is equivalent to minimizing $\Sigma \mapsto \log\det\Sigma + \mathrm{tr}(\Sigma^{-1}S)$. The difference of this expression with its value at S is equal to

$$\log\det\Sigma + \mathrm{tr}(\Sigma^{-1}S) - \log\det S - \mathrm{tr}(S^{-1}S) = -\log\det(\Sigma^{-1}S) + \mathrm{tr}(\Sigma^{-1}S) - d.$$

In terms of the eigenvalues $\lambda_1, \ldots, \lambda_d$ of the matrix $\Sigma^{-1/2}S\Sigma^{-1/2}$ this can be written as

$$-\sum_{i=1}^d \log\lambda_i + \sum_{i=1}^d \lambda_i - d = -\sum_{i=1}^d \bigl(\log\lambda_i - \lambda_i + 1\bigr).$$

Because $\log x - x + 1 \leq 0$ for all $x > 0$, with equality only for $x = 1$, this expression is nonnegative, and it is zero only if all eigenvalues are equal to one. It follows that the minimum is taken for $\Sigma^{-1/2}S\Sigma^{-1/2} = I$. ■

14.36 EXERCISE. Show that S is nonsingular with probability one if Σ is nonsingular. [Hint: show that $a^T S a > 0$ almost surely for every $a \neq 0$.]

In the genetic context we are often interested in fitting a multivariate normal distribution with a restricted parameter space. In particular, the covariance matrix is often structured through a covariance decomposition. In this case we maximize the likelihood over the appropriate subset of covariance matrices. Depending on the particular structure there may not be simple analytic formulas for the maximum likelihood estimator, but the likelihood must be maximized by a numerical routine.

As shown in the preceding proof the maximum likelihood estimator for the mean vector μ remains the sample average $\bar{X}_n$ as long as μ is a free parameter in $\mathbb{R}^n$. Furthermore, minus 2 times the log likelihood, with $\bar{X}_n$ substituted for μ is up to a constant equal to

$$\log\det(\Sigma S^{-1}) + \mathrm{tr}\bigl(\Sigma^{-1}S\bigr).$$

We may think of this expression as a measure of discrepancy between Σ and S. The maximum likelihood estimator for Σ minimizes this discrepancy, and can be viewed as the matrix in the model that is "closest" to S. This criterion for estimating the covariance matrix makes sense also without assuming normality of the observations. Moreover, rather than the sample covariance matrix S we can use another reasonable initial estimator for the covariance matrix.

In genetic applications the mean vector is typically not free, but restricted to have equal coordinates. Its maximum likelihood estimator is then often the "overall mean" of the observations. We prove this for the case of possibly non-identically distributed observations. The likelihood for μ based on independent observations

$X_1,\ldots,X_n$ with X_i possessing a $N_d(\mu 1,\Sigma_i)$-distribution is up to a constant equal to

$$\begin{aligned}\mu \mapsto -\tfrac12\sum_{i=1}^n \log\det\Sigma_i - \tfrac12\sum_{i=1}^n (X_i-\mu 1)^T\Sigma_i^{-1}(X_i-\mu 1)\\ = -\tfrac12\sum_{i=1}^n \log\det\Sigma_i - \tfrac12\operatorname{tr}\Bigl(\sum_{i=1}^n \Sigma_i^{-1}(X_i-\mu 1)(X_i-\mu 1)^T\Bigr).\end{aligned}$$

Here 1 is the vector in $\mathbb{R}^d$ with all coordinates equal to 1, so that $\mu 1=(\mu,\ldots,\mu)^T$.

14.37 Lemma. *The likelihood in the preceding display is maximized by* $\hat\mu=\sum_{i=1}^n 1^T\Sigma_i^{-1}X_i/\sum_{i=1}^n 1^T\Sigma_i^{-1}1$. *If all row sums of each matrix* Σ_i *are equal to a single constant, then* $\hat\mu=(nd)^{-1}\sum_{i=1}^n\sum_{i=1}^d X_{ij}$.

Proof. The maximum likelihood estimator minimizes the strictly convex function $\mu\mapsto\sum_{i=1}^n (X_i-\mu 1)^T\Sigma_i^{-1}(X_i-\mu 1)$. The derivative of this function is $-2\sum_{i=1}^n 1^T\Sigma^{-1}(X_i-\mu 1)$, which is zero at $\hat\mu$ as given. By convexity this is a point of minimum.

The vector of row sums of Σ_i is equal to $\Sigma_i 1$. This vector has identical coordinates equal to c if and only if $\Sigma_i 1=c1$, in which case $\Sigma_i^{-1}1=c^{-1}1$. The second assertion of the lemma follows upon substituting this in the formula for $\hat\mu$. ∎

The covariance matrices Σ_i are often indexed by a common parameter γ. Consider an observation X possessing a $N_d(\mu 1,\sigma^2\Sigma_\gamma)$-distribution, for one-dimensional parameters $\mu\in\mathbb{R}$, $\sigma^2>0$ and $\gamma\in\mathbb{R}$. The log likelihood for observing X is up to a constant equal to

$$-\tfrac12\log\sigma^2-\tfrac12\log\det\Sigma_\gamma-\tfrac12\frac{1}{\sigma^2}(X-\mu 1)^T\Sigma_\gamma^{-1}(X-\mu 1).$$

The score function is the vector of partial derivatives of this expression with respect to the parameters μ,σ^2,γ. In view of the lemma below this can be seen to be, with $\dot\Sigma_\gamma$ the derivative of the matrix Σ_γ relative to γ,

$$\dot\ell_{\mu,\sigma^2,\gamma}(x)=\begin{pmatrix}\frac{1}{\sigma^2}1^T\Sigma_\gamma^{-1}(x-\mu 1)\\ -\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}(x-\mu 1)^T\Sigma_\gamma^{-1}(x-\mu 1)\\ -\frac12\operatorname{tr}\bigl(\Sigma_\gamma^{-1}\dot\Sigma_\gamma\bigr)+\frac{1}{2\sigma^2}(x-\mu 1)^T\Sigma_\gamma^{-1}\dot\Sigma_\gamma\Sigma_\gamma^{-1}(x-\mu 1)\end{pmatrix}.$$

14.38 Lemma. *If* $t\mapsto A(t)$ *is a differentiable map from* $\mathbb{R}$ *into the invertible* $(d\times d)$-*matrices, then*
(i) $\frac{d}{dt}A(t)^{-1}=-A(t)^{-1}A'(t)A(t)^{-1}$.
(ii) $\frac{d}{dt}\det A(t)=\det A(t)\operatorname{tr}\bigl(A(t)^{-1}A'(t)^T\bigr)$.

Proof. Statement (i) follows easily from differentiating across the identity $A(t)^{-1}A(t)=I$.

For every i the determinant of an arbitrary matrix $B=(B_{ij})$ can be written $\det B=\sum_k B_{ik}\det B^{ik}(-1)^{i+k}$, for B^{ik} the matrix derived from B by deleting the

ith row and kth column. The matrix B can be thought of as consisting of the d^2 free elements B_{ij}, where the matrix B^{ik} is free of B_{ij}, for every k. It follows that the derivative of $\det B$ relative to B_{ij} is given by $\det B^{ij}(-1)^{i+j}$. Consequently, by the chain rule

$$\frac{d}{dt}\det A(t) = \sum_i \sum_j \det A(t)^{ij}(-1)^{i+j}\ A'(t)_{ij}.$$

By Cramer's formula for an inverse matrix, $(B^{-1})_{ij} = (-1)^{i+j}\det B^{ij}/\det B$, this can be written in the form of the lemma. ■

14.5 Logistic Regression

In the standard logistic regression model the observations are a random sample $(X_1, Y_1), \ldots, (X_N, Y_N)$ from the distribution of a vector (X, Y), where X ranges over $\mathbb{R}^d$ and $Y \in \{0, 1\}$ is binary, with distribution determined by

$$P(Y = 1|\, X = x) = \Psi(\alpha + \beta^T x), \qquad X \sim F.$$

Here $\Psi(x) = 1/(1+e^{-x})$ is the logistic distribution function, the intercept α is real, and the regression parameter β is a vector in $\mathbb{R}^d$. If we think of X as the covariate of a randomly chosen individual from a population and Y as his disease status, 1 referring to diseased, then the prevalence of the disease in the population is

$$p_{\alpha,\beta,F} = P(Y = 1) = \int \Psi(\alpha + \beta^T x)\, dF(x).$$

By Bayes' rule the conditional distributions of X given $Y = 0$ or $Y = 1$ are given by

$$\begin{aligned} dF_{0|\alpha,\beta,F}(x) &= \frac{\bigl(1 - \Psi(\alpha + \beta^T x)\bigr)\, dF(x)}{1 - p_{\alpha,\beta,F}}, \\ dF_{1|\alpha,\beta,F}(x) &= \frac{\Psi(\alpha + \beta^T x)\, dF(x)}{p_{\alpha,\beta,F}}. \end{aligned}$$

These conditional distributions are the relevant distributions if the data are sampled according to a case-control design, rather than sampled randomly from the population. Under the case-control design the numbers of healthy and diseased individuals to be sampled are fixed in advance, and the data consists of two random samples, from $F_{0|\alpha,\beta,F}$ and $F_{1|\alpha,\beta,F}$, respectively.

If we denote these samples by $X_1, \ldots, X_m$ and $X_{m+1}, \ldots X_{m+n}$, set $N = m+n$ and define auxiliary variables $Y_1, \ldots, Y_N$ to be equal to 0 or 1 if the corresponding X_i belongs to the first or second sample, then in both the random and the case-control design the observations can be written as $(X_1,Y_1),\ldots,(X_N,Y_N)$. The likelihoods for the two settings are

$$\begin{aligned}\mathrm{pros}(\alpha,\beta,F)&=\prod_{i=1}^N \Psi(\alpha+\beta^TX_i)^{Y_i}\bigl(1-\Psi(\alpha+\beta^TX_i)\bigr)^{1-Y_i}\,dF(X_i),\\ \mathrm{retro}(\alpha,\beta,F)&=\prod_{i=1}^m dF_{0|\alpha,\beta,F}(X_i)\prod_{i=m+1}^N dF_{1|\alpha,\beta,F}(X_i).\end{aligned}$$

The names used for these functions are abbreviations of *prospective design* and *retrospective design*, respectively, which are alternative labels for the two designs. With our notational conventions these likelihoods satisfy the relationship

$$\mathrm{pros}(\alpha,\beta,F)=\mathrm{retro}(\alpha,\beta,F)(1-p_{\alpha,\beta,F})^m p_{\alpha,\beta,F}^n.$$

Thus the two likelihoods differ by the likelihood of Bernoulli form $(1-p)^mp^n$, for $p=p_{\alpha,\beta,F}$ the prevalence of the disease. It is intuitively clear that in the case-control design this factor is not estimable, as it is not possible to estimate the prevalence $p_{\alpha,\beta,F}$. The following lemma formalizes this, and, on the positive side, shows that apart from this difference nothing is lost. In particular, the regression parameter β is typically estimable from both designs, and the profile likelihoods are proportional.

Let $\mathcal{FF}$ be the set of pairs $(F_{0|\alpha,\beta,F},F_{1|\alpha,\beta,F})$ of control-case distributions when (α,β,F) ranges over the parameter space $\mathbb{R}\times\mathbb{R}^d\times\mathcal{F}$, for $\mathcal{F}$ the set of distributions on $\mathbb{R}^d$ whose support is not contained in an affine linear space of lower dimension (i.e. if $\beta^TX=c$ almost surely for $X\sim F\in\mathcal{F}$, $\beta\in\mathbb{R}^d$ and $c\in\mathbb{R}$, then $\beta=0$ and $c=0$).

14.39 Lemma. *For all $p\in(0,1)$ and $(F_0,F_1)\in\mathcal{FF}$ there exists a unique $(\alpha,\beta,F)\in\mathbb{R}\times\mathbb{R}^d\times\mathcal{F}$ such that*

$$F_0=F_{0|\alpha,\beta,F},\qquad F_1=F_{1|\alpha,\beta,F},\qquad p=p_{\alpha,\beta,F}.$$

Furthermore, equality $F_{i|\alpha,\beta,F}=F_{i|\alpha',\beta',F'}$ for $i=0,1$ and parameter vectors $(\alpha,\beta,F),(\alpha',\beta',F')\in\mathbb{R}\times\mathbb{R}^d\times\mathcal{F}$ can happen only if $\beta=\beta'$.

Proof. For any parameter vector (α,β,F) the measures $F_{0|\alpha,\beta,F}$ and $F_{1|\alpha,\beta,F}$ are absolutely continuous, and equivalent to F. The definitions show that their density is given by

$$\log\frac{dF_{1|\alpha,\beta,F}}{dF_{0|\alpha,\beta,F}}(x)=\alpha+\beta^Tx+\log\frac{1-p_{\alpha,\beta,F}}{p_{\alpha,\beta,F}}.$$

Therefore, solving the three equations in the display of the lemma for (α,β,F) requires that $p_{\alpha,\beta,F}=p$ and

$$\alpha+\beta^Tx=\log\frac{dF_1}{dF_0}(x)-\log\frac{1-p}{p}. \tag{14.40}$$

Since $(F_0, F_1) \in \mathcal{F}\mathcal{F}$, there exists (α', β', F') such that $F_{i|\alpha',\beta',F'} = F_i$ for $i = 1, 2$, whence the right side is equal to a linear function $(\beta')^T x$ plus a constant. The equation is solved by $\beta = \beta'$ and α equal to the constant. By assumption the solution $\beta = \beta'$ is unique, and so is then the solution α.

The equations $F_{i|\alpha,\beta,F} = F_i$ for $i = 1, 2$ yield

$$\frac{1 - \Psi(\alpha + \beta^T x)}{1 - p}\, dF(x) = dF_0(x), \qquad \frac{\Psi(\alpha + \beta^T x)}{p}\, dF(x) = dF_1(x).$$

If we define F by the second equation, then $p = \int p\, dF_1 = p_{\alpha,\beta,F}$. Furthermore by a calculation using (14.40) we see that the first equation is automatically satisfied. Combining the equations we see that $F = (1-p)F_0 + pF_1$. This shows that F is indeed a probability measure, which is equivalent to F_0 and F_1.

The triple (α, β, F) thus found satisfies the three requirements, and is clearly unique. ■

The parameter p in the lemma may be interpreted as the prevalence of the disease in the full population. The lemma proves that this is not identifiable based on case-control data: for any after-the-fact value p there exist parameter values (α, β, F) that correspond to prevalence p and can produce any possible distribution for the case-control data. Fortunately, the most interesting parameter β is identifiable.

In this argument the marginal distribution F of the covariates has been left unspecified. This distribution is also not identifiable from the case-control data (it is a mixture $F = (1-p)F_0 + pF_1$ that depends on the unknown prevalence p), and it makes much sense not to model it, as it factors out of the prospective likelihood and would normally be assumed not to contain information on the regression parameter β. If we had a particular model for F (the extreme case being that F is known), then the argument does not go through. The relation $F = (1-p)F_0 + pF_1$ then contains information on F and p.

To estimate the parameter β using case-control data, we might artificially fix a value $p \in (0,1)$ and maximize the retrospective likelihood $\mathrm{retro}(\alpha, \beta, F)$ under the constraint $p_{\alpha,\beta,F} = p$. This will give an estimator $(\hat\alpha_p, \hat\beta, \hat F_p)$ of which the first and last coordinates depend on p, but with $\hat\beta$ the same for any p. Because the Bernoulli likelihood $p \mapsto (1-p)^m p^n$ is maximal for $\hat p = n/N$, the relationship between prospective and retrospective likelihoods shows that $(\hat\alpha_{\hat p}, \hat\beta, \hat F_{\hat p})$ will be the maximizer of the prospective likelihood. In particular, maximizing the prospective likelihood relative to (α, β, F) yields the correct maximum likelihood estimator for β, even if the data are obtained in a case-control design.

Similar observations apply to the likelihood ratio and score statistics in the two models.

Besides showing that these statistics are identical functions of the data, this reasoning also applies to connect the asymptotic distributions in the two models: in the retrospective model these are the same as in the prospective (i.i.d.) model with the parameters fixed to (α_p, β, F_p) solving $p_{\alpha,\beta,F} = p$, with $p = n/N$. Consider this in detail for the score statistics. Under the constraint $p_{\alpha,\beta,F} = p$ the retrospective likelihood takes the form

$$\mathrm{retro}(\alpha_p, \beta, F_p) = \prod_{i=1}^{m} \frac{\bigl(1 - \Psi(\alpha_p + \beta^T X_i)\bigr)\, dF_p(X_i)}{1-p} \prod_{i=m+1}^{N} \frac{\Psi(\alpha_p + \beta^T X_i)\, dF_p(X_i)}{p}.$$

For inference on β we drop the factors $dF_p(X_i)$, $1-p$ and p, and are left with a function of β only, which can be written in the form

$$\prod_{i=1}^{N} \bigl(1 - \Psi(\alpha_p + \beta^T X_i)\bigr)^{1-Y_i} \Psi(\alpha_p + \beta^T X_i)^{Y_i}.$$

The score statistic for testing an hypothesis on β is the derivative of the logarithm of this expression with respect to β evaluated at the maximum likelihood estimator $\hat\beta_0$ under the null hypothesis, the maximizer of the same expression over the null hypothesis. We derive the asymptotics by the method of Section 14.3. In terms of the local parameter $b = \sqrt{N}(\beta - \beta_0)$, the log likelihood is up to a constant and higher order terms equal to $b^T S_N + \frac{1}{2} b^T I_N b$, for

$$\begin{aligned} S_N &= \frac{d}{db}_{|b=0} \log \mathrm{retro}(\alpha_p, \beta_0 + b/\sqrt{N}, F_p) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \bigl(Y_i - \Psi(\alpha_p + \beta_0^T X_i)\bigr) X_i, \\ I_N &= -\frac{1}{N} \sum_{i=1}^{N} \Psi'(\alpha_p + \beta_0^T X_i) X_i X_i^T. \end{aligned}$$

(The expression for the score S_N is simplified using the identity $\Psi' = \Psi(1 - \Psi)$.) We shall show that in the retrospective model these quantities behave as the score and Fisher information in the prospective model with parameter (α_p, β_0, F_p).

In the retrospective model the observations (X_i, Y_i) form two independent samples with $Y_i = 0$ or $Y_i = 1$ and X_i distributed according to $X|\, Y = 0$ or $X|\, Y = 1$, respectively. By applying the Law of Large Numbers separately to the two samples, we see that under true parameter β_0 minus the Fisher information matrix I_N is asymptotic to

$$\begin{aligned} -\mathrm{E} I_N &= \frac{m}{N} \mathrm{E}_{\alpha_p, \beta_0, F_p} \Bigl(\Psi'(\alpha_p + \beta^T X) X X^T |\, Y = 0 \Bigr) \\ &\qquad\qquad + \frac{n}{N} \mathrm{E}_{\alpha_p, \beta_0, F_p} \Bigl(\Psi'(\alpha_p + \beta_X^T) X X^T |\, Y = 1 \Bigr) \\ &= \mathrm{E}_{\alpha_p, \beta_0, F_p} \Bigl(\Psi'(\alpha_p + \beta^T X) X X^T \Bigr), \end{aligned}$$

for $p = n/N = P_{\alpha_p, \beta_0, \mathcal{F}_p}(Y = 1)$. In the retrospective model the mean and covariance matrix of S_N are

$$\begin{aligned}\mathrm{E}S_N &= \frac{m}{\sqrt{N}}\mathrm{E}_{\alpha_p,\beta_0,F_p}\Bigl(-\Psi(\alpha_p+\beta^TX)X|\,Y=0\Bigr)\\ &\qquad\qquad + \frac{n}{\sqrt{N}}\mathrm{E}_{\alpha_p,\beta_0,F_p}\Bigl(\bigl(Y-\Psi(\alpha_p+\beta^TX)\bigr)X|\,Y=1\Bigr),\\ \mathrm{Cov}(S_N) &= \frac{m}{N}\,\mathrm{Cov}_{\alpha_p,\beta_0,F_p}\Bigl(-\Psi(\alpha_p+\beta^TX)X|\,Y=0\Bigr)\\ &\qquad\qquad + \frac{n}{N}\,\mathrm{Cov}_{\alpha_p,\beta_0,F_p}\Bigl(\bigl(Y-\Psi(\alpha_p+\beta^TX)\bigr)X|\,Y=1\Bigr).\end{aligned}$$

Arguing as for the Fisher information, we can write the expectation as $\mathrm{E}S_N = \sqrt{N}\mathrm{E}_{\alpha_p,\beta_0,F_p}\bigl(\bigl(Y-\Psi(\alpha_p+\beta^TX)\bigr)X\bigr)$, which is zero, as expected of a true score function. Similarly we can write the covariance as

$$\mathrm{E}_{\alpha_p,\beta_0,F_p}\mathrm{Cov}_{\alpha_p,\beta_0,F_p}\Bigl[\bigl(Y-\Psi(\alpha_p+\beta^TX)\bigr)X|\,Y\Bigr] = \mathrm{Cov}_{\alpha_p,\beta_0,F_p}\Bigl[\bigl(Y-\Psi(\alpha_p+\beta^TX)\bigr)X\Bigr].$$

The last equality follows from the fact that $\mathrm{E}_{\alpha_p,\beta_0,F_p}\bigl(\bigl(Y-\Psi(\alpha_p+\beta^TX)\bigr)X|\,Y\bigr) = 0$. This can be seen by writing the conditional expectations relative to $Y=0$ and $Y=1$ as integrals relative to $F_{0|\alpha_p,\beta,F_p}$ and $F_{1|\alpha_p,\beta,F_p}$, and next using the explicit forms of these distributions, the fact that $\Psi(1-\Psi)=\Psi'$ and the identity $\int\Psi'(\alpha_p+\beta^Tx)\,dF_p(x)=0$, which is a consequence of the defining identity $\int\Psi(\alpha_p+\beta^Tx)\,dF_p(x)=p_{\alpha_p,\beta,F_p}=p$, for α_p and F_p. Finally the right side of the last display can be seen to be minus $-\mathrm{E}I_N$, by calculated after conditioning on X.

As explained in Section 14.3, the score statistic for testing the local null hypothesis $H_0\colon b\in B_0$ is asymptotic to $S_N - I_N\hat b_0$, for $\hat b_0 = \mathrm{argmax}_{b\in B_0}\,\mathrm{retro}(\alpha_p,\beta_0+b/\sqrt{N},F_p)$ the maximum likelihood estimator of the local parameter under the null hypothesis. As $\hat b_0 = \mathrm{argmin}_{b\in B_0}\|I_N^{-1/2}S_N - I_N^{1/2}b\|$, this can be written in the form

$$I_N^{1/2}\bigl(I_N^{-1/2}S_N - I_N^{1/2}\hat b_0\bigr) = I_N^{1/2}\bigl(I-\Pi_{I_N^{1/2}B_0}\bigr)I_N^{-1/2}S_N.$$

Hence the asymptotics of the score statistic are completely determined by the asymptotics of S_N and I_N.

14.6 Variance Decompositions

In calculus a function $T(X_1,\ldots,X_n)$ is approximated by linear, quadratic, or higher order polynomials through a Taylor expansion. If $X_1,\ldots,X_n$ are random variables, in particular variables with a discrete distribution, then such an approximation is not very natural. Approximations by an additive function $\sum_i g_i(X_i)$, a quadratic function $\sum_{(i,j)} g_{i,j}(X_i,X_j)$ or higher order functions may still be very useful. For random variables a natural sense of approximation is in terms of variance. In this section we obtain such approximations, starting from the abstract notion of a projection.

A representation of a given random variable T as a sum of uncorrelated variables corresponds to projections of T on given subspaces of random variables. These projections are often conditional expectations. In this section we first discuss these concepts and next derive the Hoeffding decomposition, which is a general decomposition of a function of n independent variables as a sum of functions of sets of $1, 2, \ldots, n$ variables.

14.6.1 Projections

Let T and $\{S: S \in \Sigma\}$ be random variables, defined on a given probability space, with finite second moments. A random variable $\hat{S}$ is called a *projection* of T onto Σ (or L_2-projection) if $\hat{S} \in \Sigma$ and minimizes

$$S \mapsto \mathrm{E}(T - S)^2, \qquad S \in \Sigma.$$

Often Σ is a linear space in the sense that $\alpha_1 S_1 + \alpha_2 S_2$ is in Σ for every $\alpha_1, \alpha_2 \in \mathbb{R}$, whenever $S_1, S_2 \in \Sigma$. In this case $\hat{S}$ is the projection of T if and only if $T - \hat{S}$ is *orthogonal* to Σ for the inner product $\langle S_1, S_2 \rangle = \mathrm{E} S_1 S_2$. This is the content of the following theorem.

14.41 Theorem. *Let Σ be a linear space of random variables with finite second moments. Then $\hat{S}$ is the projection of T onto Σ if and only if $\hat{S} \in \Sigma$ and*

$$\mathrm{E}(T - \hat{S})S = 0, \qquad \textit{every } S \in \Sigma.$$

Every two projections of T onto Σ are almost surely equal. If the linear space Σ contains the constant variables, then $\mathrm{E}T = E\hat{S}$ and $\mathrm{cov}(T - \hat{S}, S) = 0$ for every $S \in \Sigma$.

Proof. For any S and $\hat{S}$ in Σ,

$$\mathrm{E}(T - S)^2 = \mathrm{E}(T - \hat{S})^2 + 2\mathrm{E}(T - \hat{S})(\hat{S} - S) + \mathrm{E}(\hat{S} - S)^2.$$

If $\hat{S}$ satisfies the orthogonality condition, then the middle term is zero, and we conclude that $\mathrm{E}(T - S)^2 \geq \mathrm{E}(T - \hat{S})^2$, with strict inequality unless $\mathrm{E}(\hat{S} - S)^2 = 0$. Thus, the orthogonality condition implies that $\hat{S}$ is a projection, and also that it is unique.

Conversely, for any number α,

$$\mathrm{E}(T - \hat{S} - \alpha S)^2 - \mathrm{E}(T - \hat{S})^2 = -2\alpha \mathrm{E}(T - \hat{S})S + \alpha^2 \mathrm{E} S^2.$$

If $\hat{S}$ is a projection, then this expression is nonnegative for every α. But the parabola $\alpha \mapsto \alpha^2 \mathrm{E} S^2 - 2\alpha \mathrm{E}(T - \hat{S})S$ is nonnegative if and only if the orthogonality condition $\mathrm{E}(T - \hat{S})S = 0$ is satisfied.

If the constants are in Σ, then the orthogonality condition implies $\mathrm{E}(T - \hat{S})c = 0$, whence the last assertions of the theorem follow. ■

The theorem does not assert that projections always exist. This is not true: the infimum $\inf_S \mathrm{E}(T-S)^2$ need not be achieved. A sufficient condition for existence is that Σ is closed for the second moment norm, but existence is usually more easily established directly.

The orthogonality of $T-\hat{S}$ and $\hat{S}$ yields the Pythagorean rule

$$\mathrm{E}T^2 = \mathrm{E}(T-\hat{S})^2 + \mathrm{E}\hat{S}^2.$$

If the constants are contained in Σ, then this is also true for variances instead of second moments.

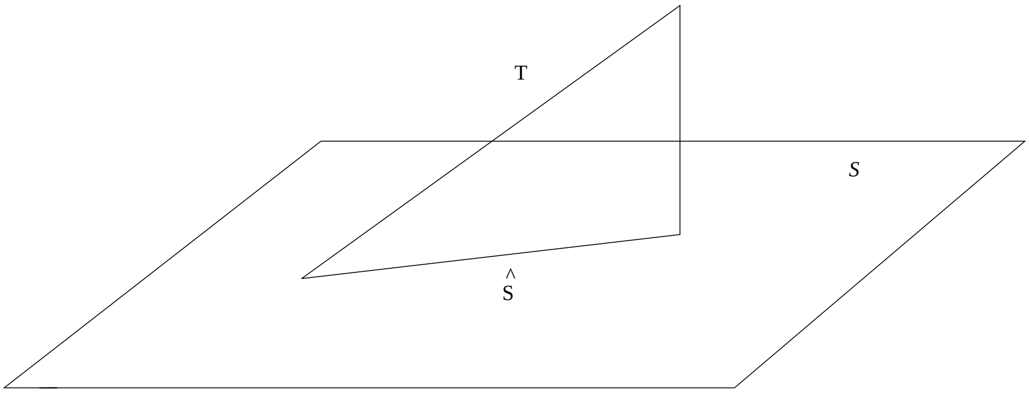

Figure 14.2. The Pythagorean rule. The vector T is projected on the linear space Σ.

The sumspace $\Sigma_1+\Sigma_2$ of two linear spaces Σ_1 and Σ_2 if random variables is the set of all variables S_1+S_2 for $S_1 \in \Sigma_1$ and $S_2 \in \Sigma_2$. The sum $\hat{S}_1+\hat{S}_2$ of the projections of a variable T onto the two subspaces is in general not the projection on the sumspace. However, this is true in the special case that the two linear spaces are orthogonal. The spaces Σ_1 and Σ_2 are called *orthogonal* if $\mathrm{E}S_1S_2=0$ for every $S_1\in\Sigma_1$ and $S_2\in\Sigma_2$.

14.42 Theorem. *If $\hat{S}_1$ and $\hat{S}_2$ are the projections of T onto orthogonal linear spaces Σ_1 and Σ_2, then $\hat{S}_1+\hat{S}_2$ is the projection onto the sumspace $\Sigma_1+\Sigma_2$.*

Proof. The variable $\hat{S}_1+\hat{S}_2$ is clearly contained in the sumspace. It suffices to verify the orthogonality relationship. Now $\mathrm{E}(T-\hat{S}_1-\hat{S}_2)(S_1+S_2) = \mathrm{E}(T-\hat{S}_1)S_1 - \mathrm{E}\hat{S}_2S_1 + \mathrm{E}(T-\hat{S}_2)S_2 - \mathrm{E}\hat{S}_1S_2$, and all four terms on the right are zero. ∎

14.43 EXERCISE. Suppose $\hat{S}_1$ and $\hat{S}_2$ are the projections of T onto linear spaces Σ_1 and Σ_2 with $\Sigma_1\subset\Sigma_2$. Show that $\hat{S}_1$ is the projection of $\hat{S}_2$ onto Σ_1.

14.6.2 Conditional Expectation

The expectation $\mathrm{E}X$ of a random variable X minimizes the quadratic form $a \mapsto \mathrm{E}(X-a)^2$ over the real numbers a. This may be expressed as: $\mathrm{E}X$ is the best "prediction" of X, given a quadratic loss function, and in the absence of additional information.

The *conditional expectation* $\mathrm{E}(X|\,Y)$ of a random variable X given a random vector Y is defined as the best "prediction" of X given knowledge of Y. Formally, $\mathrm{E}(X|\,Y)$ is a measurable function $g_0(Y)$ of Y that minimizes

$$\mathrm{E}\bigl(X - g(Y)\bigr)^2$$

over all measurable functions g. In the terminology of the preceding section, $\mathrm{E}(X|\,Y)$ is the projection of X onto the linear space of all measurable functions of Y. It follows that the conditional expectation is the unique measurable function $\mathrm{E}(X|\,Y)$ of Y that satisfies the orthogonality relation

$$\mathrm{E}\bigl(X - \mathrm{E}(X|\,Y)\bigr)g(Y) = 0, \qquad \text{every } g.$$

If $\mathrm{E}(X|\,Y) = g_0(Y)$, then it is customary to write $\mathrm{E}(X|\,Y=y)$ for $g_0(y)$. This is interpreted as the expected value of X given that $Y=y$ is observed. By Theorem 14.41 the projection is unique only up to changes on sets of probability zero. This means that the function $g_0(y)$ is unique up to sets B of values y such that $P(Y \in B) = 0$. (These could be very big sets.)

The following examples give some properties and also describe the relationship with conditional densities.

14.44 Example. The orthogonality relationship with $g \equiv 1$ yields the formula $\mathrm{E}X = \mathrm{E}\mathrm{E}(X|\,Y)$. Thus, "the expectation of a conditional expectation is the expectation". □

14.45 Example. If $X = f(Y)$ for a measurable function f, then $\mathrm{E}(X|\,Y) = X$. This follows immediately from the definition, where the minimum can be reduced to zero. The interpretation is that X is perfectly predictable given knowledge of Y. □

14.46 Example. Suppose that (X,Y) has a joint probability density $f(x,y)$ with respect to a product measure $\mu \times \nu$, and let $f(x|\,y) = f(x,y)/f_Y(y)$ be the conditional density of X given $Y=y$. Then

$$\mathrm{E}(X|\,Y) = \int x f(x|\,Y)\,d\mu(x).$$

(This is well defined only if $f_Y(y) > 0$.) Thus the conditional expectation as defined above concurs with our intuition.

The formula can be established by writing

$$\mathrm{E}\bigl(X - g(Y)\bigr)^2 = \int \Bigl[\int \bigl(x - g(y)\bigr)^2 f(x|\,y)\,d\mu(x)\Bigr]\, f_Y(y)\,d\nu(y).$$

To minimize this expression over g, it suffices to minimize the inner integral (between square brackets) by choosing the value of $g(y)$ for every y separately. For each y, the integral $\int (x-a)^2\, f(x|\,y)\, d\mu(x)$ is minimized for a equal to the mean of the density $x \mapsto f(x|\,y)$. □

14.47 Example. If X and Y are independent, then $\mathrm{E}(X|\,Y) = \mathrm{E}X$. Thus, the extra knowledge of an unrelated variable Y does not change the expectation of X.

The relationship follows from the fact that independent random variables are uncorrelated: since $\mathrm{E}(X - \mathrm{E}X)g(Y) = 0$ for all g, the orthogonality relationship holds for $g_0(Y) = \mathrm{E}X$. □

14.48 Example. If f is measurable, then $\mathrm{E}\bigl(f(Y)X|\,Y\bigr) = f(Y)\mathrm{E}(X|\,Y)$ for any X and Y. The interpretation is that, given Y, the factor $f(Y)$ behaves like a constant and can be "taken out" of the conditional expectation.

Formally, the rule can be established by checking the orthogonality relationship. For every measurable function g,

$$\mathrm{E}\bigl(f(Y)X - f(Y)\mathrm{E}(X|\,Y)\bigr)\, g(Y) = \mathrm{E}\bigl(X - \mathrm{E}(X|\,Y)\bigr)\, f(Y)g(Y) = 0,$$

because $X - \mathrm{E}(X|\,Y)$ is orthogonal to all measurable functions of Y, including those of the form $f(Y)g(Y)$. Since $f(Y)\mathrm{E}(X|\,Y)$ is a measurable function of Y, it must be equal to $\mathrm{E}\bigl(f(Y)X|\,Y\bigr)$. □

14.49 Example. If X and Y are independent, then $\mathrm{E}\bigl(f(X,Y)|\,Y = y\bigr) = \mathrm{E}f(X,y)$ for every measurable f. This rule may be remembered as follows: the known value y is substituted for Y; next, since Y carries no information concerning X, the unconditional expectation is taken with respect to X.

The rule follows from the equality

$$\mathrm{E}\bigl(f(X,Y) - g(Y)\bigr)^2 = \int\!\!\int \bigl(f(x,y) - g(y)\bigr)^2\, dP_X(x)\, dP_Y(y).$$

Once again, this is minimized over g by choosing for each y separately the value $g(y)$ to minimize the inner integral. □

14.50 Example. For any random vectors X, Y and Z,

$$\mathrm{E}\bigl(\mathrm{E}(X|\,Y,Z)|\,Y\bigr) = \mathrm{E}(X|\,Y).$$

This expresses that a projection can be carried out in steps: the projection onto a smaller set can be obtained by projecting the projection onto a bigger set a second time.

Formally, the relationship can be proved by verifying the orthogonality relationship $\mathrm{E}\bigl(\mathrm{E}(X|\,Y,Z) - \mathrm{E}(X|\,Y)\bigr)g(Y) = 0$ for all measurable functions g. By Example 14.48, the left side of this equation is equivalent to $\mathrm{E}\mathrm{E}(Xg(Y)|\,Y,Z) - \mathrm{E}\mathrm{E}(g(Y)X|\,Y) = 0$, which is true because conditional expectations retain expectations. □

14.6.3 Projection onto Sums

Let $X_1, \ldots, X_n$ be independent random vectors, and let Σ be the set of all variables of the form

$$\sum_{i=1}^n g_i(X_i),$$

for arbitrary measurable functions g_i with $\mathrm{E}g_i^2(X_i) < \infty$. The projection of a variable onto this class is known as its *Hájek projection.*

14.51 Theorem. *Let $X_1, \ldots, X_n$ be independent random vectors. Then the projection of an arbitrary random variable T with finite second moment onto the class Σ is given by*

$$\hat{S} = \sum_{i=1}^n \mathrm{E}(T|\, X_i) - (n-1)\mathrm{E}T.$$

Proof. The random variable on the right side is certainly an element of Σ. Therefore, the assertion can be verified by checking the orthogonality relation. Since the variables X_i are independent, the conditional expectation $\mathrm{E}\bigl(\mathrm{E}(T|\, X_i)|\, X_j\bigr)$ is equal to the expectation $\mathrm{E}\mathrm{E}(T|\, X_i) = \mathrm{E}T$ for every $i \neq j$. Consequently, $\mathrm{E}(\hat{S}|\, X_j) = \mathrm{E}(T|\, X_j)$ for every j, whence

$$\mathrm{E}(T - \hat{S})g_j(X_j) = \mathrm{E}\mathrm{E}(T - \hat{S}|\, X_j)g_j(X_j) = \mathrm{E}0g_j(X_j) = 0.$$

This shows that $T - \hat{S}$ is orthogonal to Σ. ∎

Consider the special case that $X_1, \ldots, X_n$ are not only independent, but also identically distributed, and that $T = T(X_1, \ldots, X_n)$ is a permutation-symmetric, measurable function of the X_i. Then

$$\mathrm{E}(T|X_i = x) = \mathrm{E}T(x, X_2, \ldots, X_n).$$

Since this does not depend on i, the projection $\hat{S}$ is also the projection of T onto the smaller set of variables of the form $\sum_{i=1}^n g(X_i)$, where g is an arbitrary measurable function.

14.6.4 Hoeffding Decomposition

The Hájek projection gives a best approximation by a sum of functions of one X_i at a time. The approximation can be improved by using sums of functions of two, or more, variables. This leads to the *Hoeffding decomposition.*

Since a projection onto a sum of orthogonal spaces is the sum of the projections onto the individual spaces, it is convenient to decompose the proposed projection space into a sum of orthogonal spaces. Given independent variables $X_1, \ldots, X_n$ and a subset $A \subset \{1, \ldots, n\}$, let H_A denote the set of all square-integrable random variables of the type

$$g_A(X_i{:}\, i \in A),$$

for measurable functions g_A of $|A|$ arguments such that

$$\mathrm{E}\bigl(g_A(X_i{:}\, i\in A)|\, X_j{:}\, j\in B\bigr)=0, \qquad \text{every } B{:}\,|B|<|A|. \tag{14.52}$$

(Define $\mathrm{E}(T|\,\emptyset)=\mathrm{E}T$.) By the independence of $X_1,\ldots,X_n$ the condition in the last display is automatically valid for any $B\subset\{1,2,\ldots,n\}$ that does not contain A. Consequently, the spaces H_A, when A ranges over all subsets of $\{1,\ldots,n\}$, are pairwise orthogonal. Stated in its present form, the condition reflects the intention to build approximations of increasing complexity by projecting a given variable in turn onto the spaces

$$[1], \qquad \Bigl[\sum_i g_{\{i\}}(X_i)\Bigr], \qquad \Bigl[\sum\sum_{i<j} g_{\{i,j\}}(X_i,X_j)\Bigr], \qquad \cdots,$$

where $g_{\{i\}}(X_i)\in H_{\{i\}}$, $g_{\{i,j\}}(X_i,X_j)\in H_{\{i,j\}}$, etcetera, and $[\cdots]$ denotes linear span. Each new space is chosen orthogonal to the preceding spaces.

Let P_AT denote the projection of T onto H_A. Then, by the orthogonality of the H_A, the projection onto the sum of the first r spaces is the sum $\sum_{|A|\le r}P_AT$ of the projections onto the individual spaces. The projection onto the sum of the first two spaces is the Hájek projection. More generally, the projections of zero, first and second order can be seen to be

$$\begin{aligned} P_\emptyset T&=\mathrm{E}T,\\ P_{\{i\}}T&=\mathrm{E}(T|\,X_i)-\mathrm{E}T,\\ P_{\{i,j\}}T&=\mathrm{E}(T|\,X_i,X_j)-\mathrm{E}(T|\,X_i)-\mathrm{E}(T|\,X_j)+\mathrm{E}T.\end{aligned}$$

Now the general formula given by the following lemma should not be surprising.

14.53 Theorem. *Let $X_1,\ldots,X_n$ be independent random variables, and let T be an arbitrary random variable with $\mathrm{E}T^2<\infty$. Then the projection of T onto H_A is given by*

$$P_AT=\sum_{B\subset A}(-1)^{|A|-|B|}\mathrm{E}(T|\,X_i{:}\, i\in B).$$

If $T\perp H_B$ for every subset $B\subset A$ of a given set A, then $\mathrm{E}(T|\,X_i{:}\, i\in A)=0$. Consequently, the sum of the spaces H_B with $B\subset A$ contains all square-integrable functions of $(X_i{:}\, i\in A)$.

Proof. Abbreviate $\mathrm{E}(T|\,X_i{:}\, i\in A)$ to $\mathrm{E}(T|\,A)$ and $g_A(X_i{:}\, i\in A)$ to g_A. By the independence of $X_1,\ldots,X_n$ it follows that $\mathrm{E}\bigl(\mathrm{E}(T|\,A)|\,B\bigr)=\mathrm{E}(T|\,A\cap B)$ for every subsets A and B of $\{1,\ldots,n\}$. Thus, for P_AT as defined in the lemma and a set C strictly contained in A,

$$\begin{aligned}\mathrm{E}(P_AT|\,C)&=\sum_{B\subset A}(-1)^{|A|-|B|}\mathrm{E}(T|\,B\cap C)\\ &=\sum_{D\subset C}\sum_{j=0}^{|A|-|C|}(-1)^{|A|-|D|-j}\binom{|A|-|C|}{j}\mathrm{E}(T|\,D).\end{aligned}$$

By the binomial formula, the inner sum is zero for every D. Thus the left side is zero. In view of the form of P_AT, it was not a loss of generality to assume that $C \subset A$. Hence P_AT is contained in H_A.

Next we verify the orthogonality relationship. For any measurable function g_A,

$$\mathrm{E}(T - P_AT)g_A = \mathrm{E}\bigl(T - \mathrm{E}(T|\,A)\bigr)g_A - \sum_{\substack{B\subset A\\ B\neq A}} (-1)^{|A|-|B|}\mathrm{E}\mathrm{E}(T|\,B)\mathrm{E}(g_A|\,B).$$

This is zero for any $g_A \in H_A$. This concludes the proof that P_AT is as given.

We prove the second assertion of the lemma by induction on $r = |A|$. If $T \perp H_\emptyset$, then $\mathrm{E}(T|\,\emptyset) = \mathrm{E}T = 0$. Thus the assertion is true for $r = 0$. Suppose that it is true for $0, \ldots, r-1$, and consider a set A of r elements. If $T \perp H_B$ for every $B \subset A$, then certainly $T \perp H_C$ for every $C \subset B$. Consequently, the induction hypothesis shows that $\mathrm{E}(T|\,B) = 0$ for every $B \subset A$ of $r-1$ or fewer elements. The formula for P_AT now shows that $P_AT = \mathrm{E}(T|\,A)$. By assumption the left side is zero. This concludes the induction argument.

The final assertion of the lemma follows if the variable $T_A\colon = T - \sum_{B\subset A} P_BT$ is zero for every T that depends on $(X_i\colon i \in A)$ only. But in this case T_A depends on $(X_i\colon i \in A)$ only and hence equals $\mathrm{E}(T_A|\,A)$, which is zero, because $T_A \perp H_B$ for every $B \subset A$. ■

14.54 EXERCISE. If $Y \in H_A$ for some nonempty set A, then $\mathrm{E}_{X_i}Y = 0$ for any i. Here E_{X_i} means: compute the expected value relative to the variable X_i, leaving all other variables X_j fixed. [Hint: $\mathrm{E}_{X_i}Y = \mathrm{E}(X_i|\,X_j\colon j \in B)$ for $B = \{1, \ldots, n\} - \{i\}$.]

14.7 EM-Algorithm

The *Expectation-Maximization Algorithm*, abbreviated *EM*, is a popular, multipurpose algorithm to compute maximum likelihood estimators in situations where the desired data is only partially observed. In many applications *missing data* models arise naturally, but the algorithm can also be applied by viewing the observed data as part of an imaginary "full observation".

We denote the observation by X, and write (X, Y) for the "full data" (X, Y), where Y may be an arbitrary random vector for which the joint distribution of (X, Y) can be defined. The probability density of the observation X can be obtained from the probability density $(x, y) \mapsto \bar p_\theta(x, y)$ of the vector (X, Y), by marginalization,

$$p_\theta(x) = \int \bar p_\theta(x, y)\, d\mu(y).$$

By definition the maximum likelihood estimator of θ based on the observation X maximizes the likelihood function $\theta \mapsto p_\theta(X)$. If the integral in the preceding display can be evaluated explicitly, then the computation of the maximum likelihood estimator becomes a standard problem, which may be solved analytically or numerically using an iterative algorithm. On the other hand, if the integral cannot be evaluated analytically, then computation of the likelihood may require numerical evaluation of an integral for every value of θ, and finding the maximum likelihood estimator may be computationally expensive. The EM-algorithm tries to overcome this difficulty by maximizing a different function.

Would the full data (X, Y) have been available, then we would have used the maximum likelihood estimator based on (X, Y). This estimator, which would typically be more accurate than the maximum likelihood estimator based on X only, is the point of maximum of the log likelihood function $\theta \mapsto \log \bar{p}_\theta(X, Y)$. We shall assume that the latter "full likelihood function" is easy to evaluate. A natural procedure if Y is not available, is to replace the full likelihood function by its conditional expectation given the observed data:

$$\theta \mapsto \mathrm{E}_{\theta_0}\bigl(\log \bar{p}_\theta(X, Y) |\, X\bigr). \tag{14.55}$$

The idea is to determine the point of maximum of this function instead of the likelihood.

Unfortunately, the expected value in (14.55) will typically depend on the parameter θ_0, a fact which has been made explicit by writing θ_0 as a subscript of the expectation operator E_{θ_0}. Because θ_0 is unknown, the function in the display cannot be used as basis of an estimation routine. The EM-algorithm overcomes this by iteration. Given a suitable first guess $\tilde{\theta}_0$ of the true value of θ, an estimator $\tilde{\theta}_1$ is determined by maximizing the criterion with $\mathrm{E}_{\tilde{\theta}_0}$ instead of E_{θ_0}. Next $\tilde{\theta}_0$ in $\mathrm{E}_{\tilde{\theta}_0}$ is replaced by $\tilde{\theta}_1$, this new criterion is maximized, etc..

```
Initialise $\tilde{\theta}_0$.
E-step: given $\tilde{\theta}_i$ compute the function $\theta \mapsto \mathrm{E}_{\tilde{\theta}_i}\bigl(\log \bar{p}_\theta(X, Y) |\, X = x\bigr)$.
M-step: define $\tilde{\theta}_{i+1}$ as the point of maximum of this function.
```

The EM-algorithm produces a sequence of values $\tilde{\theta}_0, \tilde{\theta}_1, \ldots$, and we hope that for increasing i the value $\tilde{\theta}_i$ tends to the maximum likelihood estimator.

The preceding description could suggest that the result of the EM-algorithm is a new type of estimator. This is sometimes meant to be true and then the iterations are seen as a smoothing device and stopped before convergence. However, if the algorithm is run to convergence, then the EM-algorithm is only a computational device: its iterates $\tilde{\theta}_0, \tilde{\theta}_1, \ldots$ are meant to converge to the maximum likelihood estimator.

Unfortunately, the convergence of the EM-algorithm is not guaranteed in general, although under regularity conditions it can be shown that, for every i,

$$p_{\tilde{\theta}_{i+1}}(X) \geq p_{\tilde{\theta}_i}(X). \tag{14.56}$$

Thus the EM-iterations increase the value of the likelihood, which is a good property. This does not imply convergence of the sequence $\tilde{\theta}_i$, as the sequence $\tilde{\theta}_i$ could tend to a local maximum or fluctuate between local maxima.

14.57 Lemma. *The sequence $\tilde\theta_0, \tilde\theta_1, \ldots$ generated according to the EM-algorithm yields an nondecreasing sequence of likelihoods $p_{\tilde\theta_0}(X), p_{\tilde\theta_1}(X), \ldots$.*

Proof. The density $\bar p_\theta$ of (X, Y) can be factorized as

$$\bar p_\theta(x, y) = p_\theta^{Y|X}(y|\, x) p_\theta(x).$$

The logarithm changes the product in a sum and hence

$$\mathrm{E}_{\tilde\theta_i}\bigl(\log \bar p_\theta(X, Y)|\, X\bigr) = \mathrm{E}_{\tilde\theta_i}\bigl(\log p_\theta^{Y|X}(Y|\, X)|\, X\bigr) + \log p_\theta(X).$$

Because $\tilde\theta_{i+1}$ maximizes this function over θ, this sum is bigger at $\theta = \tilde\theta_{i+1}$ than at $\theta = \tilde\theta_i$. If we can show that the first term on the right is bigger at $\theta = \tilde\theta_i$ than at $\theta = \tilde\theta_{i+1}$, then the second term must satisfy the reverse inequality, and the claim (14.56) is proved. Thus it suffices to show that

$$\mathrm{E}_{\tilde\theta_i}\bigl(\log p_{\tilde\theta_{i+1}}^{Y|X}(Y|\, X)|\, X\bigr) \le \mathrm{E}_{\tilde\theta_i}\bigl(\log p_{\tilde\theta_i}^{Y|X}(Y|\, X)|\, X\bigr).$$

This inequality is of the form $\int \log(q/p)\, dP \le 0$ for p and q the conditional densities of Y given X using the parameters $\tilde\theta_i$ and $\tilde\theta_{i+1}$, respectively. Because $\log x \le x - 1$ for every $x \ge 0$, any pair of probability densities p and q satisfies

$$\int \log(q/p)\, dP \le \int (q/p - 1)\, dP = \int_{p(x)>0} q(x)\, dx - 1 \le 0.$$

This implies the preceding display, and concludes the proof. ■

14.58 EXERCISE. Suppose that we observe a variable X which given an observable Y is normally distributed with mean Y and variance 1. Suppose that Y is normally distributed with mean θ and variance 1. Determine the iterations of the EM-algorithm, and show that the algorithm produces a sequence that converges to the maximum likelihood estimator, from any starting point.

14.8 Hidden Markov Models

Hidden Markov models are used to model phenomena in areas as diverse as speech recognition, financial risk management, the gating of ion channels or gene-finding. They are in fact Markov chain models for a given phenomenon, where the states of the chain are only partially observed or observed with error. The hidden nature of the Markov chain arises, because many systems can be thought of as evolving as a Markov process (in time or space) provided that the state space is chosen to contain enough information to ensure that the jumps of the process are indeed determined based on the current state only. This may necessitate including unobservable quantities in the states.

The popularity of hidden Markov models is also partly explained by the existence of famous algorithms to compute likelihood-based quantities. In fact, the EM-algorithm was first invented in the context of hidden Markov models for speech recognition.

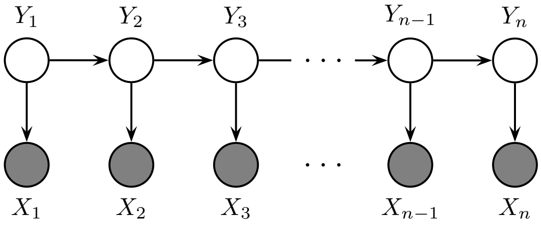

Figure 14.3. Graphical representation of a hidden Markov model. The "state variables" $Y_1, Y_2, \ldots,$ form a Markov chain, but are unobserved. The variables $X_1, X_2, \ldots$ are observable "outputs" of the chain. Arrows indicate conditional dependence relations. Given the state Y_i the variable X_i is independent of all other variables.

Figure 14.3 gives a graphical representation of a hidden Markov model. The sequence $Y_1, Y_2, \ldots,$ forms a Markov chain, and is referred to as the sequence of *state variables.* This sequence of variables is not observed ("hidden"). The variables $X_1, X_2, \ldots$ are observable, with X_i viewed as the "output" of the system at time i. Besides the Markov property of the sequence $Y_1, Y_2, \ldots$ (relative to its own history), it is assumed that given Y_i the variable X_i is conditionally independent of all other variables $(Y_1, X_1, \ldots, Y_{i-1}, X_{i-1}, Y_{i+1}, X_{i+1}, \ldots, Y_n, X_n)$. Thus the output at time i depends on the value of the state variable Y_i only.

We can describe the distribution of $Y_1, X_1, \ldots, Y_n, X_n$ completely by:

- the density π of Y_1 (giving the initial distribution of the chain).
- the set of transition densities $(y_{i-1}, y_i) \mapsto p_i(y_i|\, y_{i-1})$ of the Markov chain.
- the set of output densities $(x_i, y_i) \mapsto q_i(x_i|\, y_i)$.

The transition densities and output densities may be time-independent ($p_i = p$ and $q_i = q$ for fixed transition densities p and q), but this is not assumed. It is easy to write down the likelihood of the complete set of variables $Y_1, X_1, \ldots, Y_n, X_n$:

$$\pi(y_1)p_2(y_2|\, y_1) \times \cdots \times p_n(y_n|\, y_{n-1})\; q_1(x_1|\, y_1) \times \cdots \times q_n(x_n|\, y_n).$$

However, for likelihood inference the marginal density of the outputs $X_1, \ldots, X_n$ is the relevant density. This is obtained by integrating or summing out the hidden states $Y_1, \ldots, Y_n$. This is conceptually easy, but the n-dimensional integral may be hard to handle numerically.

The full likelihood has three components, corresponding to the initial density π, the transition densities of the chain and the output densities. In typical applications these three components are parametrized with three different parameters, which range independently. For the case of discrete state and output spaces, and under the assumption of stationarity transitions and outputs, the three components are often not modelled at all: π is an arbitrary density, and $p_i = p$ and $q_i = q$ are arbitrary transition densities. If the Markov chain is stationary in time, then the inital density π is typically a function of the transition density p.

14.8.1 Baum-Welch Algorithm

The Baum-Welch algorithm is the special case of the EM-algorithm for hidden Markov models. Historically it was the first example of an EM-algorithm.

Suppose that initial estimates $\tilde\pi$, $\tilde p_i$ and $\tilde q_i$ are given. The M-step of the EM-algorithm requires that we compute

$$\begin{aligned}
&\mathrm{E}_{\tilde\pi,\tilde p,\tilde q}\Bigl(\log \pi(Y_1)\prod_{i=2}^n p_i(Y_i|\,Y_{i-1})\prod_{i=1}^n q_i(X_i|\,Y_i)|\,X_1,\ldots,X_n\Bigr)\\
&\qquad= \mathrm{E}_{\tilde\pi,\tilde p,\tilde q}\Bigl(\log\pi(Y_1)|\,X_1,\ldots,X_n\Bigr)\\
&\qquad\qquad+\sum_{i=2}^n \mathrm{E}_{\tilde\pi,\tilde p,\tilde q}\Bigl(\log p_i(Y_i|\,Y_{i-1})|\,X_1,\ldots,X_n\Bigr)\\
&\qquad\qquad+\sum_{i=1}^n \mathrm{E}_{\tilde\pi,\tilde p,\tilde q}\Bigl(\log q_i(X_i|\,Y_i)|\,X_1,\ldots,X_n\Bigr).
\end{aligned}\tag{14.59}$$

To compute the right side we need the conditional distributions of Y_i and the pairs (Y_{i-1},Y_i) given $X_1,\ldots,X_n$ only, expressed in the initial guesses $\tilde\pi$, $\tilde p_i$ and $\tilde q_i$. It is shown below that these conditional distributions can be computed by simple recursive formulas. The computation of the distribution of Y_i given the observations is known as *smoothing*, and in the special case that $i = n$ also as *filtering*.

The E-step of the EM-algorithm requires that the right side of the preceding display be maximized over the parameters, which we may take to be π, p_i and q_i themselves provided that we remember that the parameters must be restricted to their respective parameter spaces. As this step depends on the specific models used, there is no general recipe. However, if the three types of parameters π, p_i and q_i range independently over their respective parameter spaces, then the maximization can be performed separately, using the appropriate term of the right side of (14.59) (provided the maxima are finite).

14.60 Example (Stationary transitions, nonparametric model). Assume that $p_i = p$ and $q_i = q$ for fixed but arbitrary transition densities p and q, and no other restrictions are placed on the model. The three terms on the right side of (14.59) can be written

$$\int \log\pi(y)\; p_{\tilde\pi,\tilde p,\tilde q}^{Y_1|X_1,\ldots,X_n}(y)\,d\mu(y),$$

$$\int\Bigl[\int \log p(v|\,u)\,\Bigl(\sum_{i=2}^n p_{\tilde\pi,\tilde p,\tilde q}^{Y_{i-1},Y_i|X_1,\ldots,X_n}(u,v)\Bigr)\,d\mu(v)\Bigr]\,d\mu(u),$$

$$\int \Bigl[\sum_{x\in\mathcal{X}} \log q(x|\,y)\Bigl(\sum_{i:X_i=x} p_{\tilde\pi,\tilde p,\tilde q}^{Y_i|X_1,\ldots,X_{i-1},X_i=x,X_{i+1},\ldots,X_n}(y)\Bigr)\Bigr]\,d\mu(y).$$

The first expression is the divergence between the density π and the distribution of Y_1 given $X_1,\ldots,X_n$. Without restrictions on π (other than that π is a probability density) it is maximized by taking π equal to the density of the second distribution,

$$\pi = p_{\tilde\pi,\tilde p,\tilde q}^{Y_1|X_1,\ldots,X_n}(y).$$

The inner integral (within square brackets) in the second expression is, for fixed u, the divergence between the density $v\mapsto p(v|\,u)$ and the function given by the sum (between round brackets) viewed as function of v. Thus by the same argument this expression is maximized over arbitrary transitities densities p by

$$p(v|\,u) = \frac{\sum_{i=2}^n p_{\tilde\pi,\tilde p,\tilde q}^{Y_{i-1},Y_i|X_1,\ldots,X_n}(u,v)}{\sum_{i=2}^n p_{\tilde\pi,\tilde p,\tilde q}^{Y_{i-1}|X_1,\ldots,X_n}(u)}.$$

The sum (in square brackets) in the third expression of the display can be viewed, for fixed y, also as a divergence, and hence by the same argument this term is maximized by

$$q(x|\,y) = \frac{\sum_{i:X_i=x} p_{\tilde\pi,\tilde p,\tilde q}^{Y_i|X_1,\ldots,X_{i-1},X_i=x,X_{i+1},\ldots,X_n}(y)}{\sum_{x\in\mathcal{X}}\sum_{i:X_i=x} p_{\tilde\pi,\tilde p,\tilde q}^{Y_i|X_1,\ldots,X_{i-1},X_i=x,X_{i+1},\ldots,X_n}(y)}.$$

These expressions can be evaluated using the formulas for the conditional distributions of Y_{i-1} and (Y_{i-1},Y_i) given $X_1,\ldots,X_n$, given below. □

14.8.2 Smoothing

The algorithm for obtaining formulas for the conditional distributions of the variables Y_{i-1},Y_i given the observations $X_1,\ldots,X_n$ is expressed in the functions

$$\begin{aligned}\alpha_i(y)&:= P(X_1=x_1,\ldots,X_i=x_i,Y_i=y),\\ \beta_i(y)&:= P(X_{i+1}=x_{i+1},\ldots,X_n=x_n|\,Y_i=y).\end{aligned}$$

Here the values of $x_1,\ldots,x_n$ have been omitted from the notation on the left, as these can be considered fixed at the observed values throughout. These functions can be computed recursively in i by a forward algorithm (for the α's) and a backward algorithm (for the β's), starting from the initial expressions

$$\alpha_1(y)=\pi(y)q_1(x_1|\,y), \qquad\qquad \beta_n(y)=1.$$

The *forward algorithm* is to write

$$\begin{aligned}\alpha_{i+1}(y) &= \sum_z P(x_1,\ldots,x_{i+1},Y_{i+1}=y,Y_i=z)\\ &= \sum_z q_{i+1}(x_{i+1}|\,y)p_{i+1}(y|\,z)\alpha_i(z).\end{aligned}$$

Here the argument x_i within $P(\cdots)$ is shorthand for the event $X_i = x_i$. The *backward algorithm* is given by

$$\begin{aligned}\beta_i(y) &= \sum_z P(x_{i+1},\ldots,x_n|\, Y_{i+1}=z, Y_i=y)P(Y_{i+1}=z|\, Y_i=y)\\ &= \sum_z q_{i+1}(x_{i+1}|\, z)\beta_{i+1}(z)p_{i+1}(z|\, y).\end{aligned}$$

Given the set of all α's and β's we may now obtain the likelihood of the observed data as

$$P(X_1=x_1,\ldots,X_n=x_n) = \sum_y \alpha_n(y).$$

The conditional distributions of Y_i and (Y_{i-1}, Y_i) given $X_1,\ldots,X_n$ are the joint distributions divided by this likelihood. The second joint distribution can be written

$$\begin{aligned}&P(Y_{i-1}=y, Y_i=z, x_1,\ldots,x_n)\\ &\quad= P(Y_i=z, x_i,\ldots,x_n|\, Y_{i-1}=y, x_1,\ldots,x_{i-1})P(Y_{i-1}=y, x_1,\ldots,x_{i-1})\\ &\quad= P(x_{i+1},\ldots,x_n|\, Y_i=z, x_i)P(Y_i=z, x_i|\, Y_{i-1}=y, x_1,\ldots,x_{i-1})\alpha_{i-1}(y)\\ &\quad= \beta_i(z)q_i(x_i|\, z)p_i(z|\, y)\alpha_{i-1}(y).\end{aligned}$$

By a similar argument we see that

$$\begin{aligned}P(Y_i=y, x_1,\ldots,x_n) &= P(x_{i+1},\ldots,x_n|\, Y_i=y, x_1,\ldots,x_i)P(Y_i=y, x_1,\ldots,x_i)\\ &= \beta_i(y)\alpha_i(y).\end{aligned}$$

14.61 EXERCISE. Show that $P(X_1=x_1,\ldots,X_n=x_n) = \sum_y \alpha_i(y)\beta_i(y)$ for every $i=1,\ldots,n$.

14.8.3 Viterbi Algorithm

The Viterbi algorithm computes the most likely sample path of the hidden Markov chain $Y_1,\ldots,Y_n$ given the observed outputs. It is a backward programming algorithm, also known as *dynamic programming*. The "most likely path" is the vector $(y_1,\ldots,y_n)$ that maximizes the conditional probability

$$(y_1,\ldots,y_n) \mapsto P(Y_1=y_1,\ldots,Y_n=y_n|\, X_1,\ldots,X_n).$$

This is of interest in some applications. It should be noted that there may be many possible paths, consistent with the observations, and the most likely path may be nonunique or not very likely and only slightly more likely than many other paths. Thus for many applications use of the full conditional distribution of the hidden states given the observations (obtained in the "smoothing" algorithm) is preferable.

14.9 Importance Sampling

If $Y_1, \ldots, Y_B$ is a sequence of random variables with a fixed marginal distribution P, then typically, for any sufficiently integrable function h, as $B \to \infty$,

$$\frac{1}{B}\sum_{b=1}^{B} h(Y_b) \to \int h\,dP, \qquad \text{a.s..} \tag{14.62}$$

For instance, this is true by the Law of Large Numbers if the variables Y_b are independent, but it is true under many types of dependence as well (e.g. for irreducible, aperiodic Markov chains, weakly mixing time series) In the case that the sequence $Y_1, Y_2, \ldots$ is stationary, the property is exactly described as *ergodicity*.

The convergence gives the possibility of "computing" the integral $\int h\,dP$ by generating a suitable sequence $Y_1, Y_2, \ldots$ of variables with marginal distribution P. We are then of course also interested in the speed of convergence, which roughly would be expressed by the variance of $B^{-1}\sum_{b=1}^{B} h(Y_b)$. For the variance the dependence structure of the sequence $Y_1, Y_2, \ldots$ is important, but also the "variation" of the function h. Extreme values of $h(Y_1)$ may contribute significantly to the expectation $\mathrm{E}h(Y_1)$, and even more so to the variance $\operatorname{var} h(Y_1)$. If extreme values are assumed with small probability, then we would have to generate a very long sequence $Y_1, Y_2, \ldots, Y_B$ to explore these rare values sufficiently to obtain accurate estimates. Intuitively this seems to be true whatever the dependence structure, although certain types of dependence may be of help here. One would want the sequence $Y_1, Y_2, \ldots$ to "explore" the various regions of the domain of h sufficiently well in order to obtain a good impression of the average size of h.

Importance sampling is a method to improve the Monte Carlo estimate (14.62) by generating the variables from a different distribution than P. If we are interested in computing $\int h\,dP$, but we generate $X_1, X_2, \ldots$ with marginal distribution Q, then we could use the estimate, with p and q are densities of P and Q,

$$\frac{1}{B}\sum_{b=1}^{B} h(X_b)\frac{p(X_b)}{q(X_b)},$$

Provided that $P \ll Q$ (i.e. q can be chosen positive whenever p is positive), this variable still has mean value $\int h\,dP$, and hence the same reasoning as before suggests that it can be used as an estimate of this integral for large n. The idea is to use a distribution Q for which the variance of the average in the display is small, and from which it is easy to simulate. The ideal q is proportional to the function hp, because then $h(X_b)p(X_b)/q(X_b)$ is constant and the variance 0. Unfortunately, this q is often not practical.

Consider importance sampling to calculate a likelihood in a missing data problem. The observed data is X and we wish to calculate the marginal density $p_\theta(x)$ of X under a parameter θ at the observed value x. If we can think of X as part of the full data (X, Y), then

$$L(\theta; x) = p_\theta(x) = \int p_\theta(x|\, y)\, dQ_\theta(y).$$

An importance sampling scheme to compute this expectation is to generate $Y_1, Y_2, \ldots$ from a distribution Q and estimate the preceding display by

$$\hat{L}_B(\theta; x) = \frac{1}{B}\sum_{b=1}^{B} \frac{p_\theta(x|\,Y_b)q_\theta(Y_b)}{q(Y_b)}.$$

A Monte-Carlo implementation of the method of maximum likelihood is to compute this estimate for all values of the parameter θ, and find the point of maximum of $\theta \mapsto \hat{L}_B(\theta; x)$. Alternatively, it could consist of an iterative procedure in which the iterates are estimated by the Monte-Carlo method.

The optimal distribution for importance sampling has density proportional to $q(y) \propto p_\theta(x|\,y)q_\theta(y)$, and hence is exactly the conditional law of Y given $X = x$ under θ. This is often not known, and it is also not practical to use a different distribution for each parameter θ. For rich observations x, the conditional distribution of Y given x is typically concentrated in a relatively narrow area, which may make it difficult to determine an efficient proposal density q.

* 14.10 MCMC Methods

Het principe van de methode van Bayes is eenvoudig genoeg: uitgaande van een model en een a-priori verdeling berekenen we de a-posteriori verdeling met behulp van de regel van Bayes. Het rekenwerk in de laatste stap is echter niet altijd eenvoudig. Traditioneel worden vaak a-priori verdelingen gekozen die het rekenwerk voor het gegeven model vereenvoudigen. De combinatie van de binomiale verdeling met de beta a-priori verdeling is daarvan een voorbeeld. Meer recent vervangt men het analytische rekenwerk wel door stochastische simulatie, zogenaamde *Markov Chain Monte Carlo* (of *MCMC*) methoden. In principe is het met dergelijke methoden mogelijk een willekeurige a-priori verdeling te combineren met een gegeven statistisch model. In deze subsectie geven we een zeer beknopte introductie tot deze methoden.

Gegeven een waarneming X, met realisatie x, met kansdichtheid p_θ en een a-priori dichtheid π, is de a-posteriori dichtheid proportioneel aan de functie

$$\theta \mapsto p_\theta(x)\pi(\theta).$$

In de meeste gevallen is het makkelijk om deze uitdrukking te berekenen, omdat deze functie direct gerelateerd is aan de specificatie van het statistisch model en de a-priori verdeling. Om de Bayes schatter of de a-posteriori verdeling te berekenen, is het echter nodig de integraal van de functie in het display en de integraal van θ keer de functie relatief ten opzichte van θ, voor gegeven x, te evalueren. Het feit dat dit lastig kan zijn, heeft de populariteit van Bayes schatters geen goed gedaan. Het is weinig attractief gedwongen te zijn tot een bepaalde a-priori dichtheid om wille van de eenvoud van de berekeningen.

Als de parameter θ laag-dimensionaal is, bijvoorbeeld reëelwaardig, dan is het redelijk recht-toe recht-aan om de berekeningen numeriek te implementeren, bijvoorbeeld door de integralen te benaderen met sommen. Voor hoger-dimensionale parameters, bijvoorbeeld van dimensie groter gelijk aan 4, zijn de problemen groter. Simulatie methoden hebben deze problemen sinds 1990 verzacht. MCMC methoden zijn een algemene procedure voor het simuleren van een Markov keten $Y_1, Y_2, \ldots$ waarvan de marginale verdelingen ongeveer gelijk zijn aan de a-posteriori verdeling. Voordat we de MCMC algoritmen beschrijven, bespreken we in de volgende alineas enkele essentiële begrippen uit de theorie van de Markov ketens.

Een Markov keten is een rij $Y_1, Y_2, \ldots$ stochastische grootheden waarvan de voorwaardelijke verdeling van Y_{n+1} gegeven de voorgaande grootheden $Y_1, \ldots, Y_n$ alleen van Y_n afhangt. Een equivalente formulering is dat gegeven de "huidige" variabele Y_n de "toekomstige" variabele Y_{n+1} onafhankelijk is van het "verleden" $Y_1, \ldots, Y_{n-1}$. We kunnen de variabele Y_n dan zien als de toestand op het "tijdstip" n, en voor het simuleren van de volgende toestand Y_{n+1} is het voldoende de huidige toestand Y_n te kennen, zonder interceptie van de voorgaande toestanden te kennen. We zullen alleen Markov ketens beschouwen die "tijd-homogeen" zijn. Dit wil zeggen dat de voorwaardelijke verdeling van Y_{n+1} gegeven Y_n niet afhangt van n, zodat de overgang van de ene toestand naar de volgende toestand steeds volgens hetzelfde mechanisme plaats vindt. Het gedrag van de keten wordt dan volledig bepaald door de *overgangskern* Q gegeven door

$$Q(y, B) = P(Y_{n+1} \in B|\, Y_n = y).$$

Voor een vaste y geeft $B \mapsto Q(B|\, y)$ de kansverdeling op het volgende tijdstip gegeven de huidige toestand y. Vaak wordt Q gegeven door een *overgangsdichtheid* q. Dit is de voorwaardelijke dichtheid van Y_{n+1} gegeven Y_n en voldoet aan $Q(y, B) = \int_B q(y, z)\, dz$, waarbij de integraal moet worden vervangen door een som in het discrete geval.

Een kansverdeling Π heet een *stationaire verdeling* voor de overgangskern Q als, voor iedere eventualiteit B,

$$\int Q(y, B)\, d\Pi(y) = \Pi(B).$$

Deze vergelijking zegt precies dat de stationaire verdeling behouden blijft onder de overgang van Y_n naar Y_{n+1}. Bezit Y_1 de stationaire verdeling, dan bezit ook Y_2 de stationaire verdeling, etc.. Als Q een overgangsdichtheid q bezit en Π een dichtheid π (die dan *stationaire dichtheid* wordt genoemd), dan is een equivalente vergelijking

$$\int q(y, z)\, \pi(y)\, dy = \pi(z).$$

Deze laatste vergelijking geeft een eenvoudige manier om stationaire verdelingen te karakteriseren. Een dichtheid π is een stationaire dichtheid als voldaan is aan de *detailed balance* relatie

$$\pi(y)q(y, z) = \pi(z)q(z, y).$$

Deze relatie eist dat een overgang van y naar z even waarschijnlijk is aan een overgang van z naar y, als in beide gevallen het startpunt een random punt is gekozen volgens π. Een Markov keten met deze eigenschap wordt *reversibel* genoemd. Dat de detailed balance relatie impliceert dat π een stationaire dichtheid is, kan worden gezien door de beide kanten van de relatie naar y te integreren, en gebruik te maken van de gelijkheid $\int q(z,y)\,dy = 1$, voor iedere z.

De MCMC algoritmen genereren een Markov keten met een overgangskern waarvan de stationaire dichtheid gelijk is aan de a-posteriori verdeling, met de waargenomen waarde x vast genomen. De dichtheid $y \mapsto \pi(y)$ in de voorgaande algemene discussie van Markov ketens wordt in de toepassing op het berekenen van de a-posteriori dichtheid dus vervangen door de dichtheid die proportioneel is aan $\theta \mapsto p_\theta(x)\pi(\theta)$. Gelukkig is in de simulatie schema's de proportionaliteits constante onbelangrijk.

Omdat het meestal lastig is de eerste waarde Y_1 van de keten te genereren volgens de stationaire dichtheid (= a-posteriori dichtheid) is een MCMC Markov keten meestal niet stationair. Wel convergeert de keten naar stationariteit als $n \to \infty$. In de praktijk simuleert men de keten over een groot aantal stappen, en gooit vervolgens de eerste gesimuleerde data $Y_1, \ldots, Y_b$ weg, de zogenaamde "burn-in". De resterende variabelen $Y_{b+1}, Y_{b+2}, \ldots, Y_B$ kunnen dan worden opgevat als een realisatie van een Markov keten met de a-posteriori verdeling als stationaire verdeling. Door middel van bijvoorbeeld een histogram van $Y_{b+1}, \ldots, Y_B$ verkrijgen we dan een goede indruk van de a-posteriori dichtheid, en het gemiddelde van $Y_{b+1}, \ldots, Y_B$ is een goede benadering van de Bayes schatter, de a-posteriori verwachting. De motivatie voor het gebruik van deze "empirische benaderingen" is hetzelfde als in Paragraaf , met dit verschil dat de variabelen $Y_1, Y_2, \ldots$ thans een Markov keten vormen, en dus niet onafhankelijk zijn. Voor vele Markov ketens geldt echter ook een Wet van de Grote Aantallen en deze garandeert dat ook nu gemiddelden zich asymptotisch gedragen als verwachtingen. Wel blijkt de snelheid van convergentie sterk af te hangen van de overgangskern, zodat in de praktijk het nog een hele kunst kan zijn om een MCMC algoritme op te zetten dat binnen een redelijke (CPU) tijd goede benaderingen levert.

Inmiddels bestaan vele typen MCMC algoritmen. De twee belangrijkste algoritmen, welke vaak ook samen worden gebruikt, zijn het *Metropolis-Hastings* algoritme en de *Gibbs sampler*.

14.63 Example (Metropolis-Hastings). Laat q een overgangsdichtheid waarvoor het makkelijk is om te simuleren volgens de kansdichtheid $z \mapsto q(y,z)$, voor iedere gegeven y. Definieer

$$\alpha(y,z) = \frac{\pi(z)q(z,y)}{\pi(y)q(y,z)} \wedge 1.$$

Merk op dat het voldoende is de vorm van π en q te weten; de proportionaliteits constante valt weg. Neem een vaste beginwaarde Y_0 en handel vervolgens recursief als volgt:

```
gegeven $Y_n$ genereer $Z_{n+1}$ volgens $Q(Y_n, \cdot)$.
```

```
genereer $U_{n+1}$ volgens de homogene verdeling op $[0,1]$.
if $U_{n+1} < \alpha(Y_n, Z_{n+1})$ laat $Y_{n+1} := Z_{n+1}$
else laat $Y_{n+1} := Y_n$.
```

De overgangskern P van de Markov keten $Y_1, Y_2, \ldots$ bestaat uit twee stukken, corresponderend met de "if-else" splitsing. Deze kern wordt gegeven door

$$P(y, B) = \int_B \alpha(y, z) q(y, z)\, dz + \Bigl(1 - \int \alpha(y, z) q(y, z)\, d\mu(y)\Bigr) \delta_y(B).$$

Hierin is δ_y de gedenereerde verdeling (Dirac maat) in y: gegeven $Y_n = y$ blijven we in y met kans

$$1 - \int \alpha(y, z) q(y, z)\, dz.$$

Het "andere deel" van de keten beweegt volgens de subovergangsdichtheid $\alpha(y, z) q(y, z)$. De functie α is zo gekozen dat het bereik in het interval $[0, 1]$ bevat is en zodanig dat voldaan is aan de detailed balance relatie

$$\pi(y) \alpha(y, z) q(y, z) = \pi(z) \alpha(z, y) q(z, y). \tag{14.64}$$

Dit gedeelte van de Markov keten is daarom reversibel. De beweging van y naar y van het eerste "deel" van de keten is trivialerwijze symmetrisch. Uit deze vaststellingen is gemakkelijk af te leiden dat π een stationaire dichtheid voor de Markov keten $Y_1, Y_2, \ldots$ is.

Een populaire keuze voor de overgangsdichtheid q is de *random walk kern* $q(y, z) = f(z - y)$ voor een gegeven dichtheid f. Als we f symmetrisch rond 0 kiezen, dan reduceert $\alpha(y, z)$ tot $\pi(z)/\pi(y)$. De keuze van een goede kern is echter niet eenvoudig. Het algemene principe is een overgangskern q te kiezen die "bewegingen" naar variabelen Z_{n+1} in de gehele drager van π voorstelt in de eerste stap van het algoritme, en tegelijkertijd niet te vaak tot de "else" stap leidt, omdat dit de efficiëntie van het algoritme nadelig zou beïnvloeden. In MCMC jargon heet het dat we een overgangskern q zoeken die "voldoende mixing is", "voldoende de ruimte afzoekt", en "niet te vaak blijft hangen". □

14.65 Example (Gibbs Sampler). De Gibbs sampler reduceert het probleem van simuleren uit een hoog-dimensionale a-posteriori dichtheid tot herhaald simuleren uit lager-dimensionale verdelingen. Het algoritme wordt vaak gebruikt in combinatie met de Metropolis-Hastings sampler, als geen geschikte overgangsdichtheid q voor de Metropolis-Hastings algoritme voor handen is.

Veronderstel dat π een dichtheid is afhankelijk van m variabelen, en veronderstel dat we over een procedure beschikken om variabelen te genereren uit ieder van de voorwaardelijke dichtheden

$$\pi_i(x_i |\, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_m) = \frac{\pi(x)}{\int \pi(x)\, d\mu_i(x_i)}.$$

Kies een gegeven beginwaarde $Y_0 = (Y_{0,1}, \ldots, Y_{0,m})$, en handel vervolgens recursief op de volgende wijze:

```
Gegeven $Y_n = (Y_{n,1}, \ldots, Y_{n,m})$,
genereer $Y_{n+1,1}$ volgens $\pi_1(\cdot \,|\, Y_{n,2}, \ldots, Y_{n,m})$.
genereer $Y_{n+1,2}$ volgens $\pi_2(\cdot \,|\, Y_{n+1,1}, Y_{n,3} \ldots, Y_{n,m})$
                                   $\vdots$
genereer $Y_{n+1,m}$ volgens $\pi_m(\cdot \,|\, Y_{n+1,1}, \ldots, Y_{n+1,m-1})$.
```

De coördinaten worden dus om de beurt vervangen door een nieuwe waarde, steeds conditionerend op de laatst beschikbare waarde van de andere coördinaten. Men kan nagaan dat de dichtheid π stationair is voor ieder van de afzonderlijke stappen van het algoritme. □

14.66 Example (Ontbrekende data). Veronderstel dat in plaats van "volledige data" (X, Y) we slechts de data X waarnemen. Als $(x, y) \mapsto p_\theta(x, y)$ een kansdichtheid van (X, Y) is, dan is $x \mapsto \int p_\theta(x, y)\, dy$ een kansdichtheid van de waarneming X. Gegeven een a-priori dichtheid π is de a-posteriori dichtheid derhalve proportioneel aan

$$\theta \mapsto \int p_\theta(x, y)\, d\mu(y)\, \pi(\theta).$$

We kunnen de voorgaande MCMC algoritmen toepassen op deze a-posteriori dichtheid. Als de marginale dichtheid van X (de integraal in het voorgaande display) echter niet analytisch kan worden berekend, dan is het lastig om de MCMC schema's te implementeren.

Een alternatief is om de marginale verdeling niet te berekenen, en de niet-waargenomen waarden Y mee te simuleren. In de Bayesiaanse notatie is de a-posteriori verdeling de voorwaardelijke verdeling van een denkbeeldige variabele $\bar\Theta$ gegeven de waarneming X. Dit is de marginale verdeling van de voorwaardelijke verdeling van het paar $(\bar\Theta, Y)$ gegeven X. Als we in staat zouden zijn een rij variabelen $(\bar\Theta_1, Y_1), \ldots, (\bar\Theta_n, Y_n)$ volgens de laatste voorwaardelijke verdeling te genereren, dan zouden de eerste coördinaten $\bar\Theta_1, \ldots, \bar\Theta_n$ van deze rij trekkingen uit de a-posteriori verdeling zijn. Marginalizeren van een empirische verdeling is hetzelfde als "vergeten" van sommige variabelen, en dit is computationeel heel gemakkelijk!

Dus kunnen we een MCMC algoritme toepassen om variabelen $(\bar\Theta_i, Y_i)$ te simuleren uit de kansdichtheid die proportioneel is aan de afbeelding $(\theta, y) \mapsto p_\theta(x, y)\pi(\theta)$, met x gelijk aan de waargenomen waarde van de waarneming. Vervolgens gooien we de Y-waarden weg. □

14.11 Gaussian Processes

A *Gaussian process* $(X_t: t \in T)$ indexed by an arbitrary set T is a collection of random variables X_t defined on a common probability space such that the *finite-dimensional marginals*, the stochastic vectors $(X_{t_1}, \ldots, X_{t_n})$ for finite set s $t_1, \ldots, t_n \in T$, are multivariate-normally distributed. Because the multivariate normal distribution is determined by its mean vector and covariance matrix, the marginal distributions of a Gaussian process are determined by the *mean function* and *covariance function*

$$t \mapsto \mu(t) = \mathrm{E} X_t, \qquad (s,t) \mapsto C(s,t) := \mathrm{cov}(X_s, X_t).$$

A covariance function is symmetric in its arguments, and it is *nonnegative-definite* in the sense that for every finite set $t_1, \ldots, t_n$ the $(n \times n)$-matrix $\bigl(S(t_i, t_j)\bigr)$ is nonnegative-definite. By *Kolmogorov's extension theorem* any function μ and symmetric, nonnegative function C are the mean and covariance function of some Gaussian process.

The mean and covariance function also determine the distribution of countable sets of variables X_t, but not the complete *sample paths* $t \mapsto X_t$. This is usually solved by working with a "regular" version of the process, such as a version with continuous or right-continuous sample paths. While a Gaussian process as a collection of variables exists for every mean function and every nonnegative-definite covariance function, existence of a version with such regular sample paths is not guaranteed and generally requires a nontrivial proof.

A Gaussian process $(X_t: t \in \mathbb{R})$ indexed by the reals is called *stationary* if the distribution of $(X_{t_1+h}, \ldots, X_{t_n+h})$ is the same for every h, and $t_1, \ldots, t_n$. This is equivalent to the variables having the same mean and the covariance function being a function of the difference $s - t$: for some constant μ, some function $C: \mathbb{R} \to \mathbb{R}$ and every s, t,

$$\mu = \mathrm{E} X_t, \qquad \mathrm{cov}(X_s, X_t) = C(s-t).$$

By the symmetry of covariance, the function C is necessarily symmetric about 0. The variance $\sigma^2 = \mathrm{var}\, X_t = C(0)$ is of course also constant, and $\mathrm{E}(X_s - X_t)^2 = 2C(0) - 2C(s-t)$ is a function of $|s-t|$. Conversely a Gaussian process with constant mean, constant variance σ^2 and $\mathrm{E}(X_s - X_t)^2 = 2B\bigl(|s-t|\bigr)$ for some function B is stationary, with $C(t) = \sigma^2 - B\bigl(|t|\bigr)$.

An *Ornstein-Uhlenbeck process* is the stationary Gaussian process with mean zero, and covariance function

$$\mathrm{E} X_s X_t = e^{-|s-t|}.$$

14.12 Renewal Processes

A *renewal process* is a *point process* $T_0 = 0 < T_1 < T_2 < \cdots$ on the positive real line, given by the cumulative sums $T_n = X_1 + X_2 + \cdots + X_n$ of a sequence of independent, identically distributed, positive random variables $X_1, X_2, \ldots$. A *delayed renewal process* satisfies the same definition, except that X_1 is allowed a different distribution than $X_2, X_3, \ldots$. Alternatively, the term "renewal process" is used for the corresponding number of "renewals" $N_t = \max\{n\colon T_n \leq t\}$ up till time t, or for the process $\bigl(N(B)\colon B \in \mathcal{B}\bigr)$ of counts $N(B) = \#\{n\colon T_n \in B\}$ of the number of events falling in sets B belonging to some class $\mathcal{B}$ (for instance, all intervals or all Borel sets).

In the following we consider the delayed renewal process. We write F_1 for the distribution of X_1 and F for the common distribution function of $X_2, X_3, \ldots$. The special case of a renewal process corresponds to $F_1 = F$. For simplicity we assume that F is not a lattice distribution, and possesses a finite mean μ.

The nth renewal time T_n has distribution $F_n = F_1 * F^{(n-1)*}$, for $*$ denoting convolution and F^{k*} the convolution of k copies of F. The *renewal function* $m(t) = \mathrm{E}N_t$ gives the mean number of renewals up till time t. By writing $N_t = \sum_{n=1}^{\infty} 1_{T_n \leq t}$ we see that

$$m(t) = \sum_{n=1}^{\infty} F_1 * F^{(n-1)*}(t).$$

The quotient N_t/t is the number of renewals per time instant. For large t this variable and its mean are approximately equal to $1/\mu$: as $t \to \infty$,

$$\begin{aligned} \frac{N_t}{t} &\to \frac{1}{\mu}, \qquad \text{a.s.,} \\ \frac{m(t)}{t} &\to \frac{1}{\mu}. \end{aligned} \tag{14.67}$$

Roughly speaking, this means that there are h/μ renewals in a long interval of length h. For distant intervals this is also true for short intervals. By the (elementary) *renewal theorem* the expected number of renewals $m(t+h) - m(t)$ in the interval $(t, t+h]$ satisfies, for all $h > 0$, as $t \to \infty$,

$$m(t+h) - m(t) \to \frac{h}{\mu}.$$

For this last result it is important that F is not a lattice distribution.

By the definitions T_{N_t} is the last event in $[0, t]$ and T_{N_t+1} is the first event in $(t, \infty]$. The *excess time* or *residual life* at time t is the variable $E_t = T_{N_t+1} - t$, giving the time to the next renewal. In general, the distribution of E_t is dependent on t. However, for $t \to \infty$,

$$P(E_t \leq y) \to \frac{1}{\mu} \int_0^y \bigl(1 - F(x)\bigr)\, dx \colon= F_\infty(y).$$

The distribution F_∞, with density $x \mapsto \bigl(1 - F(x)\bigr)/\mu$, is known as the *stationary distribution.*

This name is explained by the fact that the delayed renewal process with initial distribution F_1 equal to the stationary distribution F_∞ is stationary, in the sense that the distribution of the counts $N(B+t)$ in a set B shifted by t is the same as the distribution of $N(B)$, for every t.[†] In fact, given the distribution F of $X_2, X_3, \ldots$, the stationary distribution is the unique distribution making the process stationary. For a stationary renewal process the distribution of the residual life time E_t is equal to F_∞ for every t. This shows that for every fixed t, the future points (in (t,∞)) arrive in intervals that have the same distribution as the intervals $E_1, E_2, \ldots$ in the original point process starting from 0. Because for a stationary process the increments $m(t+h) - m(t)$ depend on h only, the asymptotic relationships $m(t+h) - m(t) \to h/\mu$ and $m(t)/t \to 1/\mu$ become equalities for every t.

Given a renewal process $0 < T_1 < T_2 < \cdots$ we can define another point process $0 < S_1 < S_2 < \cdots$ by *random thinning*: one keeps or deletes every T_i with probabilities p and $1-p$, respectively, and defines $S_1 < S_2 < \cdots$ as the remaining time points. A randomly thinned delayed renewal process is again a delayed renewal process, with initial and renewal distributions given by

$$G_1(y) = \sum_{r=1}^\infty p(1-p)^{r-1} F_1 * F^{(r-1)*}(y),$$
$$G(y) = \sum_{r=1}^\infty p(1-p)^{r-1} F^{r*}(y).$$

It can be checked from these formulas that a randomly thinned stationary renewal process is stationary, as is intuitively clear.

14.12.1 Poisson Process

The *Poisson process* is the renewal process with $F_1 = F$ the exponential distribution (with mean μ; intensity $1/\mu$). It has many special properties:

(i) ((*Lask of memory.*) The residual life E_t is distributed as the renewal times X_i, for every t; $F_\infty = F$.

(ii) The process $\bigl(N(B)\colon B \in \mathcal{B}\bigr)$ is stationary.

(iii) The distribution of $N(B)$ is Poisson with mean $\lambda(B)/\mu$.

(iv) For pairwise disjoint sets $B_1, \ldots, B_k$, the variables $N(B_1), \ldots, N(B_k)$ are independent.

(v) The process $(N_t\colon t \geq 0)$ is Markov.

(vi) Given the number of events $N(B)$ in an interval B, the events are distributed uniformly over B (i.e. given by the ordered values of a random sample of size $N(B)$ from the uniform distribution on B).

(vii) A random thinning with retention probability p yields a Poisson process with intensity p/μ. The Poisson process is the only renewal process which after random thinning possesses renewal distribution G that belongs to the scale family of the original renewal distribution F.

[†] In terms of the process $(N_t\colon t \geq 0)$ stationarity is the same as *stationary increments*: the distribution of $N_{t+h} - N_t$ depends on $h > 0$ only and not on $t \geq 0$.

* **14.12.2 Proofs.** For proofs of the preceding and more, see Grimmett-Stirzaker, 1992, Probability and Random Processes, Chapter 10, or Karlin and Taylor, 1975, A first Course in Stochastic Processes, Chapter 5, or the following concise notes.

The more involved results are actually consequences of two basic results. For two functions $A, B\colon [0,\infty) \to \mathbb{R}$ of bounded variation let $A * B$ be the function $A * B(t) = \int_0^t A(t-x)\, dB(x) = \int_0^t B(t-x)\, dA(x)$.

The first basic result is, that, for every given bounded function a and given distribution function F on $(0,\infty)$,

$$A = a + A * F \quad \& \quad A \in \text{LB} \qquad \Longleftrightarrow \qquad A = a + \tilde{m} * a. \tag{14.68}$$

Here $\tilde{m} = \sum_{n=1}^\infty F^{n*}$ is the renewal function corresponding to F, and $A \in \text{LB}$ means that the function A is bounded on bounded intervals.[‡]

The second, much more involved result is the *renewal theorem*, which says that the function $A = a + \tilde{m} * a$ for a given "directly Riemann integrable" function a, satisfies

$$\lim_{t\to\infty} A(t) \to \frac{1}{\mu} \int_0^\infty a(x)\, dx.$$

Linear combinations of monotone, integrable functions are examples of directly Riemann integrable functions.[♭]

By conditioning on the time of the first event the renewal function of a delayed renewal process can be seen to satisfy

$$m = F_1 + \tilde{m} * F_1. \tag{14.69}$$

[♯] In view of (14.68) the function m also satisfies the *renewal equation* $m = F_1 + m * F$.

By conditioning on the time of the first event the residual time probability $A_y(t) := P(E_t > y)$ can be shown to satisfy $A_y = a_y + \tilde{A}_y * F_1$, for $a_y(t) = 1 - F_1(t+y)$ and $\tilde{A}_y(t) = P(\tilde{E}_t > y)$ the residual life of the renewal process without delay.[†] In particular, for the renewal process without delay we have $\tilde{A}_y = \tilde{a}_y + \tilde{A}_y * F$, for $\tilde{a}_y(t) = 1 - F(t+y)$. In view of (14.68) this implies that $\tilde{A}_y = \tilde{a}_y + \tilde{m} * \tilde{a}_y$. Substituting this in the equation for A_y, we see that $A_y = a_y + \tilde{a}_y * F_1 + \tilde{m} * \tilde{a}_y * F_1 = a_y + \tilde{a}_y * m$, by (14.69).

The function $A = a + \tilde{m} * a$ corresponding to $a = 1_{[0,h]}$ satisfies $A(t+h) = \tilde{m}(t+h) - \tilde{m}(t)$. Therefore, the renewal theorem gives that $\tilde{m}(t+h) - \tilde{m}(t) \to \int_0^\infty a(x)\, dx = h/\mu$, as $t \to \infty$.[‡] For the delayed renewal process the equation (14.68) allows to write $m(t+h) - m(t)$

[‡] The proof of this result starts by showing that $\tilde{m} = F + \tilde{m} * F$, which is the special case of (14.69) with $F_1 = F$ and $m = \tilde{m}$ (below). Next if A is given by the equation on the right, then $A * F = a * F + a * \tilde{m} * F = a * F + a * (\tilde{m} - F) = a * \tilde{m} = A - a$, and hence A is a solution of the equation on the left, which can be shown to be locally bounded. The converse implication follows if $A = a + \tilde{m} * a$ is the only locally bounded solution. The difference D of two solutions satisfies $D = D * F$. By iteration this yields $D = D * F^{*n}$ and hence $|D(t)| \leq F^{*n}(t) \sup_{0 \leq s \leq t} |D(s)|$, which tends to 0 as $n \to \infty$.

[♭] See W. Feller, (1971). An introduction to probability theory and its applications, volume II, page 363.

[♯] Write $m(t) = \text{EE}(N_t|\, X_1)$ and note that $\text{E}(N_t|\, X_1 = x)$ is equal to 0 if $t < x$ and equal to 1 plus the expected number of renewals at $t - x$ in a renewal process without delay.

[†] Note that $P(E_t > y|\, X_1 = x)$ is equal to 1 if $x > t + y$ equal to 0 if $t < x \leq t + y$ and equal to $\tilde{A}_y(t-x)$ if $x \leq t$.

[‡] Actually Feller derives the general form of the renewal theorem from this special case.

as

$$F_1(t+h) - F_1(t) + \int_0^t \bigl(\tilde m(t+h-x) - \tilde m(t-x)\bigr)\, dF_1(x) + \int_t^{t+h} \tilde m(t+h-x)\, dF_1(x).$$

The integrand in the third integral is bounded by $\max_{0\le t\le h} \tilde m(t) < \infty$, and hence the integral tends to zero as $t \to \infty$, as does the first term. Because the integrand in the middle integral tends pointwise to h/μ and is bounded, the integral tends to $\int h/\mu\, dF_1 = h/\mu$ as $t \to \infty$ by the dominated convergence theorem. This extends the renewal theorem to the delayed renewal process.

An application of the renewal theorem to the equation $\tilde A_y = \tilde a_y + \tilde A_y * F$, immediately yields that $P(\tilde E_t > y) = \tilde A_y(t) \to \mu^{-1} \int_0^\infty \bigl(1 - F(x+y)\bigr)\, dx = 1 - F_\infty(y)$. The equation $A_y = a_y + \tilde A_y * F_1$, where $a_y(t) = 1 - F_1(t+y) \to 0$ as $t \to \infty$, allows to extend this to the delayed renewal process: $A_y(t) \to 1 - F_\infty(y)$.

If the delayed renewal process is stationary, then $m(s+t) = m(s) + m(t)$, whence m is a linear function. Substitution of $m(t) = ct$ in the renewal equation $m = F_1 + m * F$ readily yields that $F_1 = F_\infty$. Substitution of $m(t) = t/\mu$ and $F_1 = F_\infty$ into the equation $A_y = a_y + \tilde a_y * m$ yields after some algebra that $A_y = 1 - F_\infty$. This shows that the residual life distribution is independent of t and equal to F_1, so that the process "starts anew" at every time point.

The proofs of the two statements (14.67) are based on the inequalities $T_{N_t} \le t \le T_{N_t+1}$. The first statement follows by dividing these inequalities by N_t and noting that T_{N_t}/N_t and T_{N_t+1}/N_t tend to μ by the strong law of large numbers applied to the variables X_i, as $N_t \to \infty$ almost surely. For the second statement one establishes that $N_t + 1$ is a stopping time for the filtration generated by $X_1, X_2, \ldots$, for every fixed t, so that, by Wald's equation,

$$\mathrm{E} T_{N_t+1} = \mathrm{E} \sum_{i=1}^{N_t+1} X_i = \mathrm{E}(N_t+1)\mu.$$

Combination with the inequality $T_{N_t+1} \ge t$ immediately gives $\liminf \mathrm{E} N_t/t \ge 1/\mu$. If the X_i are uniformly bounded by a constant c, then also $\mathrm{E} T_{N_t+1} \le \mathrm{E} T_{N_t} + c \le t + c$, which implies that $\limsup \mathrm{E} N_t/t \le 1/\mu$. Given general X_i, we have $N_t \le N_t^c$ for N^c the renewal process corresponding to the truncated variables $X_i \wedge c$. By the preceding $\limsup \mathrm{E} N_t/t \le 1/\mu^c$, where $\mu^c = \mathrm{E} X_1 \wedge c \uparrow \mu$, as $c \to \infty$.

14.13 Markov Processes

A continuous time *Markov process*[♭] on a countable state space $\mathcal{X}$ is a stochastic process $(X_t\colon t \ge 0)$ such that, for every $0 \le t_1 < t_2 < \cdots < t_n < \infty$ and $x_1, \ldots, x_n \in \mathcal{X}$,

$$P(X_{t_n} = x_n |\, X_{t_{n-1}} = x_{n-1}, \ldots X_{t_0} = x_0) = P(X_{t_n} = x_n |\, X_{t_{n-1}} = x_{n-1}).$$

The Markov process is called *homogeneous* (or said to have *stationary transitions*) if the right side of the preceding display depends on the time instants only through

[♭] Markov processes on countable state space are also called *Markov chains.*

their difference $t_n - t_{n-1}$. Equivalently, there exists a matrix-valued function $t \mapsto P_t$ such that, for every $s < t$ and $x, y \in \mathcal{X}$,

$$P(X_t = y|\, X_s = x) = P_{t-s}(x, y).$$

The *transition matrices* P_t are (square) stochastic matrices of dimension the cardinality of $\mathcal{X}$.[♯] The collection of matrices $(P_t\colon t \geq 0)$ form a semigroup (i.e. $P_s P_t = P_{s+t}$ and $P_0 = I$) for matrix multiplication, by the *Chapman-Kolmogorov equations*

$$P(X_{s+t} = y|\, X_0 = x) = \sum_z P(X_{s+t} = y|\, X_s = z) P(X_s = z|\, X_0 = x).$$

The *generator* of the semigroup is its derivative at zero

$$A = \frac{d}{dt}_{|t=0} P_t = \lim_{t \downarrow 0} \frac{P_t - I}{t}.$$

This derivative can be shown to exist, entrywise, as soon as the semigroup is standard. A semigroup $(P_t\colon t \geq 0)$ is called *standard* if the map $t \mapsto P_t(x, y)$ is continuous at 0 (where $P_0 = I$ is the identity), for every $x, y \in \mathcal{X}$. It is immediate from its definition that the diagonal elements of a generator are nonpositive and its off-diagonal nonnegative. It can also be shown that its row sums satisfy $\sum_y A(x, y) \leq 0$, for every x. However, in general the diagonal elements can be equal to $-\infty$, and the row sums can be strictly negative. The semigroup is called *conservative* if the generator A is finite everywhere and has row sums equal to 0.

We also call a matrix A *conservative* if it is satisfies, for every $x \neq y$,

$$-\infty < A(x, x) \leq 0, \qquad A(x, y) \geq 0, \qquad \sum_y A(x, y) = 0.$$

We shall see below that every such matrix is the generator of a semigroup of a Markov chain. Conservatism is a natural restriction, although not every generator is conservative. It excludes so-called instantaneous states x, defined as states with $A(x, x) = -\infty$.[†]

The semigroup is called *uniform* if the maps $t \mapsto P_t(x, y)$ are continuous at 0 uniformly in its entries (x, y). This strengthening of standardness can be shown to be equivalent to finiteness and uniform boundedness of the diagonal of the generator A [G& R, 6.10.5]. Any uniform semigroup is conservative. We shall also call a conservative matrix *uniform* if $\sup_x A(x, x) > -\infty$.

Continuity of the semigroup is equivalent to $P(X_t = x|\, X_0 = x) \to 1$ as $t \downarrow 0$, for every $x \in \mathcal{X}$, and is a mild requirement, which excludes only some pathological Markov processes. For instance, it is satisfied if every sample path $t \mapsto X_t$ is right continuous (relative to the discrete topology on the state space, thus meaning that the process remains for a short while in every state x it reaches). Uniform

[♯] "Stochastic means $P_t(x, y) \geq 0$ for every x, y and $\Sigma_y P_t(x, y) = 1$ for every x.

[†] There are examples of standard Markov chains with $A(x, x) = -\infty$ for every x.

continuity strengthens the requirement to uniformity in $x \in \mathcal{X}$, and does exclude some processes of interest. For instance a "pure birth" process on the state space $\mathbb{N}$, whose states are the numbers of individuals present and where the birth rate is proportional to this number. On the other hand, continuous Markov semigroups on finite state spaces are (of course) automatically uniformly continuous.

The *Kolmogorov backward equation* and *Kolmogorov forward equation* are, for $t \geq 0$,

$$\frac{d}{dt}P_t = AP_t = P_tA.$$

The backward equation is satisfied by any conservative semigroup, and both equations are valid for uniform semigroups. The equations are often used in the reverse direction by starting with a generator A, and next trying to solve the equations for a semigroup $(P_t\colon t \geq 0)$ under the side conditions that P_t is a stochastic matrix for every t and $P_0 = I$. This is always possible for a uniform generator, with the unique solution given as $P_t = e^{tA}$.[‡] A solution exists also for a conservative generator, but the corresponding Markov chain may "explode" in finite time.

A continuous time Markov process $(X_t\colon t \geq 0)$ with right continuous paths gives rise to an accompanying *jump chain*, defined as the sequence of consecutive states visited by the chain, and a sequence of *holding times*, consisting of the intervals between its jump times. Together the jump chain and holding times give a complete description of the sample paths, at least up till the time of *first explosion.* True explosion is said to occur if the Markov process makes infinitely many jumps during a finite time interval; the first explosion time is then defined as the sum of the (infinitely many) holding times. True explosion cannot happen if the Markov semigroup is uniformly continuous, and the first explosion time is then defined as infinity.

The behaviour of a conservative Markov process (until explosion) has a simple description: after the chain reaches a state x, it will remain there for an exponentially distributed holding time with mean $1/\bigl|A(x,x)\bigr|$, and will next jump to another state $y \neq x$ with probability $A(x,y)/\bigl|A(x,x)\bigr|$. A more precise description is that the distribution of the chain (as given by the semigroup) is the same as the distribution of a Markov chain such that

(i) The jump chain $Y_0, Y_1, \ldots$ is a discrete time Markov chain on $\mathcal{X}$ with transition matrix Q given by $Q(x,y) = A(x,y)/\bigl|A(x,x)\bigr|1_{x\neq y}$ for every $x, y \in \mathcal{X}$.

(ii) Given $Y_0, \ldots, Y_{n-1}$ the first n holding times are independent exponentially distributed variables with parameters $-A(Y_0,Y_0), \ldots, -A(Y_{n-1},Y_{n-1})$.

Thus the diagonal elements of the generator matrix A determine the mean waiting times, and the off-diagonal elements in a given row are proportional to the transition probabilities of the jump chain.

If X_0 possesses a distribution π, then X_t possesses the distribution πP_t given

[‡] The exponential of a matrix is defined by its power series: $e^A = \Sigma_{n=0}^\infty A^n/n!$, with $A^0 = I$, where the convergence of the series can be interpreted entrywise if the state space is finite and relative to an operator norm otherwise.

by

$$(\pi P_t)(y) = \sum_x \pi(x) P_t(x,y).$$

(The notation πP_t is logical if π is viewed as a horizontal vector $\bigl(\pi(x){:}\, x \in \mathcal{X}\bigr)$ that premultiplies the matrix P_t.) A probability distribution π on the state space is called a *stationary distribution* if $\pi P_t = \pi$ for every t. If the semigroup is uniform, then this is equivalent to the equation $\pi A = 0$.

A Markov process is called *irreducible* if $P_t(x,y) > 0$ for any pair of states $x,y \in \mathcal{X}$ and some $t > 0$. (In that case $P_t(x,y) > 0$ for all $t > 0$ if the semigroup is standard.) For an irreducible Markov process with standard semigroup $P_t(x,y) \to \pi(x)$ as $t \to \infty$, for every $x,y \in \mathcal{X}$, as soon as there exists a stationary distribution π, and $P_t(x,y) \to 0$ otherwise. In particular, there exists at most one stationary distribution. For a finite state space there always exists a stationary probability distribution.

* **14.13.1 Proofs.** In this section we give full proofs for the case that the state space is finite and some other insightful proofs. For the general case, see K.L.Chung, Markov chains with stationary transition probabilities, or D. Freedman, Markov Chains.

The proof of existence of a generator can be based on the identify $(P_h - I)(\sum_{j=0}^{n-1} P_{jh}) = P_{nh} - I$. Provided that the second matrix on the left is invertible, this identity can be written as

$$\frac{P_h - I}{h} = (P_{nh} - I)\Bigl(h\sum_{j=1}^{n-1} P_{jh}\Bigr)^{-1}.$$

For a standard semigroup on a finite space the matrix $h\sum_{j=1}^{n-1} P_{jh}$ tends to $\int_0^t P_s\, ds$ if $h = t/n$ and $n \to \infty$, which itself approaches the identity as $t \downarrow 0$ and hence indeed is nonsingular if t is sufficiently close to 0. Thus the right side tends to a limit.

This argument can be extended to uniform semigroups on countable state spaces by interpreting the matrix-convergence in the operator sense. Every P_t is an operator on the space ℓ_∞ of bounded sequences $x = (x_1, x_2, \ldots)$, normed by $\|x\|_\infty = \sup_i |x_i|$. The norm of an operator $P{:}\, \ell_\infty \to \ell_\infty$ is $\|P\| = \sup_x \sum_y \bigl|P(x,y)\bigr|$. A semigroup $(P_t{:}\, t \geq 0)$ is uniform exactly when $P_t \to I$ in the operator norm, as $t \downarrow 0$, because $\|P_t - I\| = 2\sup_x \bigl(1 - P_t(x,x)\bigr)$.

It follows also that for a uniform semigroup the convergence $(P_t - I)/t \to A$ is also relative to the operator norm. This implies immediately that $\|A\| < \infty$, whence $\sup_x A(x,x) > -\infty$, and $\sum_y A(x,y) = 0$, for every x, because $(P_t - I)/t$ has these properties for every $t \geq 0$. The norm convergence also allows to conclude that the limit as $h \to 0$ of $(P_{t+h} - P_t)/h = (P_h - I)P_t/h = P_t(P_h - I)/h$ exists in norm sense for every $t \geq 0$, and is equal to $AP_t = P_tA =$. This establishes the Kolmogorov backward and forward equations.

The backward equation can be established more generally for conservative semigroups by noting that, for any finite set $\mathcal{Z} \subset \mathcal{X}$ that does not contain x,

$$\begin{aligned}\sum_{z\in\mathcal{Z}}\bigl(P_h(x,z)-I(x,z)\bigr)P_t(z,y)/h &\to \sum_{z\in\mathcal{Z}}A(x,z)P_t(z,y),\\ \sum_{z\notin\mathcal{Z}}\bigl|P_h(x,z)-I(x,z)\bigr|P_t(z,y)/h &\le \sum_{z\notin\mathcal{Z}}P_h(x,z)/h=\bigl(1-\sum_{z\in\mathcal{Z}}P_t(x,z)\bigr)/h\\ &\to -\sum_{z\in\mathcal{Z}}A(x,z)=\sum_{z\notin\mathcal{Z}}A(x,z).\end{aligned}$$

If $\mathcal{Z}$ increases to $\mathcal{X}$, then the right side of the first equation converges to $\sum_z A(x,z)P_t(x,z)$ and the right side of the last equation to zero.

For a uniform generator A the matrix exponential e^{tA} is well defined, and solves the Kolmogorov backward and forward equations. To see that it is the only solution we iterate the backward equation to see that the nth derivative satisfies $P_0^{(n)} = A^n$. By an (operator-valued) Taylor expansion it follows that $P_t = \sum(t^n/n!)A^n = e^{tA}$.

To prove that a Markov chain with uniform semigroup satisfies the description in terms of jump chain and holding times, it now suffices to construct such a chain and show that it has generator A.

Given an arbitrary stochastic matrix R of dimension the cardinality of $\mathcal{X}$ consider the stochastic process X with state space $\mathcal{X}$ which starts arbitrarily at time 0, and changes states only at the times of a Poisson process $(N_t\colon t \ge 0)$ with intensity λ, when it moves from a current state x to another state y with probability $R(x,y)$. One can check that this process X is Markovian and has stationary transitions, given by

$$P_t(x,y)=\sum_{n=0}^{\infty}e^{-\lambda t}\frac{(\lambda t)^n}{n!}R^n(x,y)=e^{\lambda t(R-I)}(x,y).$$

(Here $R^n(x,y)$ is the (x,y)-entry of the matrix R^n, the n-fold product of the matrix R with itself, which is the n-step transition matrix of the chain.) It follows that X has generator $A=\lambda(R-I)$.

If the matrix R has zero diagonal, then each jump time of N is also a jump time of X and hence is exponentially distributed with mean λ. More generally, we allow the diagonal of R to be positive and then a move of X at a jump time of N may consist of a "jump" from a state x to itself. The waiting time for a true jump from x to another state is longer than t if all jumps that occur before t are to x itself. Because the possibility that N has n jumps is Poisson with mean λt, this event has probability

$$\sum_{n=0}^{\infty}\frac{e^{-\lambda t}(\lambda t)^n}{n!}R(x,x)^n=e^{-\lambda\bigl(1-R(x,x)\bigr)t}.$$

Thus the waiting time from x to another state is exponentially distributed with intensity $\lambda\bigl(1-R(x,x)\bigr)=-A(x,x)$. By construction, from state x the process jumps to y with probability $R(x,y)$, and if jumps to x itself are discarded the process jumps to $y \neq x$ with probability $R(x,y)/\sum_z R(x,z) = A(x,y)/A(x,x)$. It follows that X is a Markov process with generator A with the desired jump chain and holding times.

Given a uniform generator A, the matrix $R = I + \lambda^{-1}A$ for some fixed $\lambda \ge \sup_x |A(x,x)|$ is a stochastic matrix, and can be used in the preceding. The resulting generator $\lambda(R-I)$ is exactly A.

That a stationary distribution of a uniform Markov chain π satisfies $\pi A = 0$ is immediate from the definition of A and the fact πP_t is constant. Conversely, by the backward equation $\pi A = 0$ implies that πP_t has derivative zero and hence πP_t is constant in t.

* 14.13.2 Kronecker Product

The *Kronecker product* of a $(k\times l)$-matrix $A = (A_{ij})$ and a $(m\times n)$-matrix $B = (b_{ij})$ is the $(km \times ln)$-matrix

$$A\otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1l}B \\ a_{21}B & a_{22}B & \cdots & a_{2l}B \\ \vdots & \vdots & & \vdots \\ a_{k1}B & a_{k2}B & \cdots & a_{kl}B \end{pmatrix}.$$

If (X_n) and (Y_n) are Markov chains with state spaces $\mathcal{X} = \{x_1,\ldots,x_k\}$ and $\mathcal{Y} = \{y_1,\ldots,y_m\}$, then $\bigl((X_n,Y_n)\bigr)$ is a stochastic process with state space $\mathcal{X}\times\mathcal{Y}$. If the two Markov chains are independent, then the joint chain is also a Markov chain. If $\mathbb{P} = (p_{ij})$ and $\mathbb{Q} = (q_{ij})$ are the transition matrices of the chains, then $\mathbb{P}\otimes\mathbb{Q}$ is the transition matrix of the joint chain, if the space $\mathcal{X}\times\mathcal{Y}$ is ordered as $(x_1,y_1),(x_1,y_2),\ldots,(x_1,y_m),(x_2,y_1),\ldots,(x_2,y_m),\ldots,(x_k,y_1),\ldots,(x_k,y_m)$.

* 14.14 Multiple Testing

A test is usually designed to have probabilities of errors of the first kind smaller than a given "level" α. If N tests are carried out at the same time, each with a given level α_j, for $j = 1,\ldots,N$, then the probability that one or more of the tests takes the wrong decision is obviously bigger than the level of each separate test. The worst case scenario is that the critical regions of the tests are disjoint, so that the probability that some test takes the wrong decision is equal to the sum $\sum_j \alpha_j$ of the error probabilties of the individual tests. Motivated by this worst case the *Bonferroni correction* simply decreases the level of each individual test to $\alpha_j = \alpha/N$, so that the overall level is certainly bounded by $\sum_j \alpha_j \le N(\alpha/N) = \alpha$. However, usually the critical regions of the tests do overlap and the overall probability of an error of the first kind is much smaller than this upper bound. The question is then how to design a "less conservative" correction for multiple testing.

This question is not easy to answer in general, as it depends on the joint distribution of the test statistics. If the jth test rejects the jth null hypothesis $H_0^j\colon \theta\in\Theta_0^j$ if the observation X falls in a critical region K^j, then an error of the first kind relative to H_0^j occurs if $X\in K^j$ and $\theta\in\Theta_0^j$. The multiple testing procedure rejects all null hypotheses H_0^j such that $X\in K^j$. The actual state of affairs may be that the null hypotheses H_0^j for every j in some set $J\subset\{1,\ldots,N\}$ are true, i.e. the true parameter θ is contained in $\cap_{j\in J}\Theta_0^j$. Some error of the first kind then occurs if $X\in\cup_{j\in J}K^j$ and hence the overall probability of an error of the first kind is

$$\sup_{\theta\in\cap_{j\in J}\Theta_0^j} P_\theta\bigl(X\in\cup_{j\in J}K^j\bigr). \tag{14.70}$$

The multiple testing procedure is said to provide *strong control of the familywise error rate* if this expression is smaller than a prescribed level α, for any possible configuration J of true hypotheses. Generally strong control is desirable, but one also defines *weak control* as the property that the expression in the display is smaller than α if all null hypothesis are true (i.e. in the case that $J=\{1,\ldots,N\}$).

The overall error (14.70) can be bounded by the errors of the individual test by

$$\sup_{\theta\in\cap_{j\in J}\Theta_0^j}\sum_{j\in J} P_\theta\bigl(X\in K^j\bigr)\le \sum_{j\in J}\sup_{\theta\in\Theta_0^j} P_\theta\bigl(X\in K^j\bigr).$$

If all individual tests have level α, then the right side is bounded by $\#J\,\alpha\le N\alpha$. This proves that the Bonferroni correction gives strong control. It also shows why the Bonferroni correction is conservative: not only is the sum-bound pessimistic, because the critical regions $\{X\in K^j\}$ may overlap, also the final bound $N\alpha$ is based on the possibility that all null $\#J$ hypotheses are correct.

Interestingly, if the tests are in reality stochastically independent, then the union bound is not bad. Under independence,

$$P_\theta\bigl(X\in\cup_{j\in J}K^j\bigr)=1-\prod_{j\in J}\bigl(1-P_\theta(X\in K^j)\bigr).$$

If all tests are of level α and $\theta\in\cap_{j\in J}\Theta_0^j$, then the right side is bounded by $1-(1-\alpha)^{\#J}$. This is of (of course) smaller than the Bonferroni bound $\#J\alpha$, but it is not much smaller. To obtain overall level α_0, the Bonferroni correction would suggest to use size $\alpha_0/\#J$, while the preceding display suggests the value $1-(1-\alpha_0)^{1/\#J}$. The quotient of these values tends to $-\log(1-\alpha_0)/\alpha_0$ if $\#J\to\infty$, which is approximately 1.025866 for $\alpha_0=0.05$.

However, positive dependence among the hypotheses is more common than independence. Because there are so many different forms of (positive) dependence, there is no general recipe to handle this situation. If the critical regions take the form $K^j=\{T^j>c\}$ for some test statistics $T^j=T^j(X)$ and critical value c, then the event that some hypothesis H_0^j for $j\in J$ is rejected is

$$\bigl\{\max_{j\in J}T^j>c\bigr\}.$$

The overall error probability then requires the distribution of the maximum of the test statistics, which is a complicated function of their joint distribution.

If this joint distribution is not analytically available, then *permutation or randomization methods* may help out. These are based on the assumption that the data X can be split into two parts $X=(V,W)$, where under the full null hypothesis W possesses a fixed distribution. For instance, in the two-sample problem, where X

consists of the observations for both samples, the vector V is defined as the unordered set of observed values stripped from the information to which sample they belong and W is defined as the sample labels; under the null hypothesis that the two samples arise from the same distribution any assignment W of the values to the two samples is equally likely. The idea is to compare the observed value of a test statistic to the set of values obtained by randomizing W, but keeping V fixed.

We denote this randomization by $\tilde{W}$; mathematically this should be a random variable defined on the same probability space as $X = (V, W)$, so that we can speak of the joint distribution of $(V, W, \tilde{W})$. It is convenient to describe the permutation procedure through a corrected p-value. The *p-value* for the jth null hypothesis H_0^j is a random function $P_j(X)$ of $X = (V, W)$, where it is understood that H_0^j is rejected if $P_j(X)$ is smaller than a prescribed level. The *Westfall and Young single step method* adapts these p-values for multiple testing by replacing $P_j(X)$ by

$$\tilde{P}_j(X) = P\Bigl(\min_{i=1,\ldots,N} P_i(V, \tilde{W}) \le P_j(V, W) |\, X\Bigr).$$

The probability is computed given the value of X and hence refers only to the randomization variable $\tilde{W}$. In practice, this probability is approximated by simulating many copies of $\tilde{W}$ and calculating the fraction of copies such that $\min_{i=1,\ldots,N} P_i(V, \tilde{W})$ is smaller than $P_j(V, W)$.

14.71 Theorem. *Suppose that the variables W and $\tilde{W}$ are conditionally independent given V. If the conditional distributions of the random vectors $\bigl(P_j(V, \tilde{W})\colon j \in J\bigr)$ and $\bigl(P_j(V, W)\colon j \in J\bigr)$ given V are the same under $\cap_{j\in J} H_0^j$, then under $\cap_{j\in J} H_0^j$,*

$$P\Bigl(\exists j \in J\colon \tilde{P}_j(X) < \alpha\Bigr) \le \alpha.$$

Consequently, if the condition holds for every $J \subset \{1, \ldots, N\}$, then the Westfall and Young procedure gives strong control over the familywise error rate.

Proof. Let $F^{-1}(\alpha |\, X)$ be the α-quantile of the conditional distribution of the variable $\min_i P_i(V, \tilde{W})$ given $X = (V, W)$. By the first assumption of the theorem, this quantile is actually a function of V only. By the definition of $\tilde{P}_j(X)$ it follows that $\tilde{P}_j(X) < \alpha$ if and only if $P_j(X) < F^{-1}(\alpha |\, X)$. Hence

$$\begin{aligned} P\Bigl(\exists j \in J\colon \tilde{p}_j(V, W) < \alpha |\, V\Bigr) &= P\Bigl(\min_{j\in J} p_j(V, W) < F^{-1}(\alpha |\, X) |\, V\Bigr) \\ &= P\Bigl(\min_{j\in J} p_j(V, \tilde{W}) < F^{-1}(\alpha |\, X) |\, V\Bigr). \end{aligned}$$

In the last step we use that $F^{-1}(\alpha |\, X)$ is deterministic given V, and the assumed equality in conditional distribution. The right side becomes bigger if we replace the minimum over J by the minimum over all indices $\{1, \ldots, N\}$, which yields $F\bigl(F^{-1}(\alpha |\, X) - |\, X\bigr) \le \alpha$. The proof of the first assertion follows by taking the expecation over V. ∎

The condition that W and $\tilde{W}$ are conditionally independent given V is explicitly stated, but ought to be true by construction. It is certainly satisfied if $\tilde{W}$ is produced by an "external" randomization device.

The second condition of the theorem ought also be satisfied by construction, but it is a bit more subtle, because it refers to the set of true hypotheses. If the conditional distributions of W and $\tilde{W}$ given V are the same under $\cap_{j\in J}H_0^j$, then certainly the conditional distributions of the p-values are the same. However, this assumption is overly strong. Imagine carrying out two two-sample tests based on two measurements (say of characteristics A and B) on each individual in groups of cases and controls. The variable V can then be taken to be a matrix of two rows whose columns are bivariate vectors recording the measured characteristics A and B on the combined cases and controls stripped from the case/control status. The variable W can be taken equal to a binary vector recording for each column of V whether it corresponds to a case or a control, and we construct $\tilde{W}$ as a random permutation of W. If cases and controls are indistinguishable for both characteristics (i.e. the full null hypothesis is true), then the values V are not informative on W, and hence W and $\tilde{W}$ are equal in conditional distribution given V. However, if the cases and controls are indistinguishable with respect to A (i.e. H_0^A holds), but differ in characteristic B, then V *is* informative on W, and hence the conditional distributions of W and $\tilde{W}$ given V differ. On the other hand, under H_0^A the first coordinates of the set of bivariate vectors V are an i.i.d. sample and hence if we (re)order these first coordinates using $\tilde{W}$, we obtain the same distribution as before. As long as the p-values $P_A(X)$ for testing characteristic A depend only on these first coordinates (and there is no reason why they would not), the condition that $p_A(V,W)$ and $p_A(V,\tilde{W})$ are equally distributed given V is satisfied.

Note that in this example the units that are permuted are the vectors of all measurements on a single individual. This is because we also want the joint (conditional) distributions of $\bigl(p_j(X){:}\, j\in J\bigr)$ to remain the same after permutation.

14.14.1 False Discovery Rate

If interest is in very many hypotheses (e.g. $N\geq 1000$), then controlling the level is perhaps not useful, as this is tied to preventing even a single error of the first kind. Instead we might accept that a small number of true null hypotheses is rejected provided that this is a small fraction of all hypotheses that are rejected. The *false discovery rate* (FDR) formalized this as the expected quotient

$$FDR(\theta)=\mathrm{E}_\theta\frac{\#\{j{:}\, X\in K^j,\theta\in\Theta_0^j\}}{\#\{j{:}\, X\in K^j\}}.$$

An FDR of at most 5% is considered to be a reasonable criterion.

The following procedure, due to *Benjamini and Hochberg* (BH), is often applied. The procedure is formulated in terms of the p-values P_j of the N tests N.

(i) Place the p-values in increasing order: $P_{(1)}\leq P_{(2)}\leq\cdots\leq P_{(N)}$, and let the hypothesis $H_0^{(j)}$ correspond to $P_{(j)}$.

(ii) Reject all null hypotheses $H_0^{(j)}$ with $NP_{(j)}\leq j\alpha$.

(iii) Reject in addition all null hypotheses with p-value smaller than one of the rejected hypotheses in (ii).

The Benjamini-Hochberg method always rejects more hypotheses than the Bonferroni method, as the latter rejects an hypotheses if $NP_j \leq \alpha$, whereas the Benjamini-Hochberg method employs an extra factor j in the evaluation of the jth p-value (in (ii)). However, the procedure controls the FDR at level α, or nearly so. Below we prove that

$$FDR(\theta; BH) \leq \frac{\#\{j\colon \theta \in \Theta_0^j\}}{N}\,\alpha\,(1+\log N). \tag{14.72}$$

Thus by decreasing α by the factor $1+\log N$ the Benjamini-Hochberg method gives control of the FDR at level α. This factor is modest as compared to the factor N used by the Bonferroni correction. Moreover, the theorem below also shows that the factor can be deleted if the tests are stochastically independent or are positively dependent (in a particular sense).

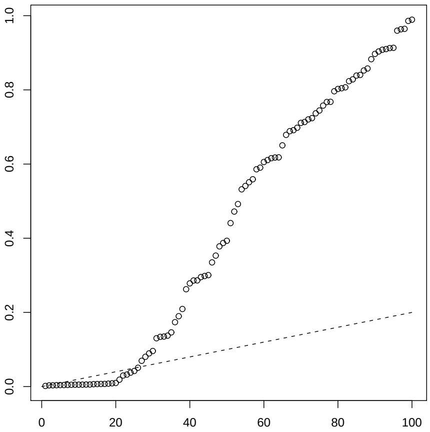

Figure 14.4. Illustration of the Benjamimi-Hochberg procedure for multiple testing. The points are the order p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(100)}$ (vertical axis) plotted agains the numbers $1, 2, \ldots, 100$ (horizontal axis). The dotted curve is the line $p \mapsto 0.20/100p$. The hypotheses corresponding to p-values left of the intersection of the two curves are rejected at level $\alpha = 0.20$. (In case of multiple intersections, we would use the one that is most to the right.)

The factor $\#\{j\colon \theta \in \Theta_0^j\}/N$ is the fraction correct null hypotheses. If this fraction is far from unity, then the Benjamini-Hochberg procedure will be conservative. One may attempt to estimate this fraction from the data and use the estimate to increase the value of α used. In the case of independent tests it works to replace α by, for any given $\lambda \in (0,1)$,

$$\alpha\frac{(1-\lambda)N}{\#\{j\colon P_j > \lambda\}+1}, \qquad \lambda \in (0,1). \tag{14.73}$$

Unfortunately, this "adaptive" extension of the Benjamini-Hochberg procedure appears to perform less well in the case the tests are dependent.

In the following theorem we assume that the P_j are random variables with values in $[0,1]$ whose distribution under the null hypothesis $\theta\in\Theta_0^j$ is stochastically larger than the uniform distribution, i.e. $P_\theta(P_j\le x)\le x$ for every $x\in[0,1]$. This expresses that they are true p-values, in that the test which rejects if $P_j\le\alpha$ is of level $P_\theta(P_j\le\alpha)\le\alpha$ for every $\alpha\in(0,1)$.

14.74 Theorem. *If P_j is stochastically larger than the uniform distribution under every $\theta\in\Theta_0^j$, then (14.72) holds. If, moreover, $P_1,\ldots,P_N$ are independent or the function $x\mapsto P_\theta\bigl(K(P_1,\ldots,P_N)\ge y|\,P_j=x\bigr)$ is decreasing for every $\theta\in\Theta_0^j$ and every coordinate-wise decreasing function $K\colon[0,1]^N\to\mathbb{N}$, then also*

$$FDR(\theta;BH)\le\frac{\#\{j\colon\theta\in\Theta_0^j\}}{N}\alpha, \tag{14.75}$$

Finally, if α in the BH-procedure is replaced by (14.73) and $P_1,\ldots,P_N$ are independent, then this remains true.

Proof. Let $P=(P_1,\ldots,P_N)$ be the vector of p-values and let $K(P)=\max\{j\colon NP_{(j)}\le j\alpha\}$. The definition (i)-(iii) of the Benjamini-Hochberg procedure shows that $H_0^{(j)}$ is rejected if and only if $j\le K(P)$, or equivalently $NP_{(j)}\le K(P)$. In other words, the hypothesis H_0^j is rejected if and only if $NP_j\le K(P)\alpha$. The FDR can therefore be written as

$$\mathrm{E}_\theta\frac{\#\{j\colon P_j\le K(P)\alpha/N,\theta\in\Theta_0^j\}}{K(P)}=\sum_{j:\theta\in\Theta_0^j}\mathrm{E}_\theta\Bigl(\frac{1\{P_j\le K(P)\alpha/N\}}{K(P)}\Bigr).$$

The sum is smaller than its number of terms times the maximal by $(\alpha/N)(1+\log N)$, to prove (14.72), and by α/N, to prove (14.75). Because the expectation is taken under $\theta\in\Theta_0^j$, the variables P_j are stochastically larger than the uniform distribution.

The desired inequality for (14.72) therefore follows immediately from the first assertion of Lemma 14.76 below.

The function K is a coordinate-wise decreasing function of $P_1,\ldots,P_N$. For independent $P_1,\ldots,P_N$ the map $x\mapsto P_\theta\bigl(K(P)\ge y|\,P_j=x\bigr)$ is therefore decreasing. In the remaining case this is true by assumption. The desired inequality to prove (14.75) therefore also follows from Lemma 14.76.

To prove the last assertion of the theorem we set $G(P)=(1-\lambda)N/\bigl(\#\{j\colon P_j>\lambda\}+1\bigr)$ and redefine $K(P)$ as $K(P)=\max\{j\colon NP_{(j)}\le j\alpha G(P)\}$. We repeat the first part of the proof to see that the (adaptive) FDR can be written as

$$\sum_{j:\theta\in\Theta_0^j}\mathrm{E}_\theta\Bigl(\frac{1\{P_j\le K(P)\alpha/NG(P)\}}{K(P)}\Bigr).$$

If P^j is the vector P with the jth coordinate P_j replaced by 0, then $G(P^j) \geq G(P)$. Hence the preceding display is smaller than

$$\sum_{j:\theta\in\Theta_0^j} \mathrm{E}_\theta\Bigl(\frac{1\{P_j \leq K(P)\alpha/NG(P^j)\}}{K(P)}\Bigr) \leq \sum_{j:\theta\in\Theta_0^j} \mathrm{E}_\theta \frac{\alpha}{N} G(P^j),$$

by Lemma 14.76, applied conditionally given $(P_i{:}\, j \neq j)$. The variable $G(P^j)$ is bounded above by $(1-\lambda)N/\bigl(\#\{i \neq j{:}\, \theta \in \Theta_o^i, P_i > \lambda\} + 1\bigr)$, in which each of the P_i is subuniform, so that the variable $G(P^j)$ is stochastically bounded above by $(1-\lambda)N/(1+B^j)$ for B^j binomially distributed with parameters $\#\{i \neq j{:}\, \theta \in \Theta_o^i\}$ and $1-\lambda$. ∎

14.76 Lemma. *Let (P, K) be a an arbitrary random vector with values in $[0,1] \times \{1,2,\ldots,N\}$. If P is stochastically larger than the uniform distirbution, then, for every $c \in (0,1)$,*

$$\mathrm{E}\Bigl(\frac{1\{P \leq cK\}}{K}\Bigr) \leq c\bigl(1 + \log(c^{-1} \wedge N)\bigr).$$

If the function $x \mapsto P(K \geq y |\, P = x)$ is decreasing for every y, then this inequality is also true without the factor $1 + \log(c^{-1} \wedge N)$.

Proof. The left side of the lemma can be written in the form

$$\begin{aligned}\mathrm{E}\int_K^\infty \frac{1}{s^2}\,ds 1\{P \leq cK\} &= \int_0^\infty \mathrm{E}1\{K \leq s, P \leq cK\}\frac{ds}{s^2}\\ &\leq \int_0^\infty \mathrm{E}1\{P \leq c\lfloor s\rfloor \wedge cN\}\frac{ds}{s^2} \leq \int_0^\infty \bigl(c\lfloor s\rfloor \wedge cN \wedge 1\bigr)\,\frac{ds}{s^2}.\end{aligned}$$

Here $\lfloor s\rfloor$ is biggest integer not bigger than s, and it is used that K is integer-valued. The last expression can be computed to be equal to $c(1/2+1/3+\cdots+1/D)+(cN\wedge 1)/D$, for D the smallest integer bigger than or equal to $c^{-1} \wedge N$. This expression is bounded by $c\bigl(1+\log(c^{-1}\wedge N)\bigr)$. This completes the proof of the first assertion.

By assumption the conditional distribution of K given $P = x$ is stochastically decreasing in x. This implies that the corresponding quantile functions $u \mapsto Q(u|\, x)$ decrease as well: $Q(u|\, x') \leq Q(u|\, x)$ if $x' \geq x$, for every $u \in [0,1]$.

Fix $u \in (0,1)$. The function $x \mapsto cQ(u|\, x) - x$ assumes the value $cQ(u|\, 0) \geq 0$ at $x = 0$ and is strictly decreasing on $[0,1]$. Let x^* be the unique point where the function crosses the horizontal axis, or be equal to $x^* = 0$ or $x^* = 1$ if the function is never positive or always positive, respectively. In all cases $cQ(u|\, P) \geq cQ(u|\, x^*-) \geq x^*$ als $P < x^*$ and the event $\{P \leq cQ(u|\, P)\}$ is contained in the event $\{P \leq x^*\}$. It follows that

$$\mathrm{E}\Bigl(\frac{1\{P \leq cQ(u|\, P)\}}{Q(u|\, P)}\Bigr) \leq \mathrm{E}\Bigl(\frac{1\{P \leq x^*\}}{Q(u|\, P)}\Bigr) \leq \mathrm{E}\Bigl(\frac{1\{P \leq x^*\}}{x^*/c}\Bigr) \leq c.$$

This is true for every $u \in (0,1)$ and hence also for U replaced by a uniform random variable that is independent of P. Because the variable $Q(U|\, x)$ is distributed according to the conditional distribution of K given $P = x$, the vector $\bigl(P, Q(U|\, P)\bigr)$ is distributed as the vector (P, K). Thus we obtain the assertion of the lemma. ∎