

Inference of Genetic Relationship

Deena C. Mobbs

Ph.D.

University of Edinburgh

1993



Abstract

Interpretation of population structure is a major objective in many aspects of population biology. The structure of a population is defined by the pattern of relationships between the individuals within it. Recent advances in molecular techniques have made it possible to identify individuals uniquely, and to infer the genealogical relationships among members of a group; these techniques can be used for many different species. Most research into population structure using such molecular techniques involves comparison of offspring with one or both putative parents. For example, when reconstructing genealogies, the first step is often to find possible offspring-parent links or offspring/parent-pair triplets. This, however, is only possible when adequate data on members of the parental generation are available for analysis. In natural populations, such as *Drosophila*, the parents will not be present when the offspring emerge while in many other species full sibships may not be easily identifiable by observation of behaviour. Direct comparison of genotypes between two or more generations is not always feasible.

The question that has been addressed in this thesis is whether useful information on genealogy or family size distribution can be obtained from analysis of data gathered from members of a single generation. In order to investigate this problem, two different analysis methods have been developed and applied to both simulated and field genotype data.

The likelihood approach investigates the likelihood for different relationships among the members of a sample of individuals from the population being studied. This method is based on the probability of obtaining the genotypes observed in the sample, given that the hypothetical relationship under consideration is the true one. For this a small number of named individuals are considered simultaneously. Results show that the relationship between these individuals can be estimated, and that the accuracy of the estimator depends on the true relationship; the probability that the relationship will be correctly inferred is greater when the individuals are more closely related.

The analysis of variance method uses a comparison of within and between group variances to estimate mean family size. Specific individuals are not identified in this approach. Results suggest that some inferences can be made regarding population structure, but that the precision of these inferences depends on the true structure in the population.

The analysis techniques have been applied to both simulated and field experimental data for a range of sample sizes and relationships. A single locus with two alleles provides the minimum level of information and the accuracy of the inferences made regarding genealogy are compared with those obtained using increasing numbers of alleles and loci. When data are available from only one generation, it has been shown that these analysis methods enable certain inferences to be made regarding the genealogical relationship within a sample of individuals drawn from the study population.

Acknowledgements

I would like to thank the following for their help in the preparation of this thesis.

- My supervisors, Professor Bill Hill and Professor Eric Renshaw, for their comments, discussions and advice; with particular thanks to Bill who suggested this topic of research with the memorable words “Well, suppose I gave you a banana that had flies crawling out of it ...”.
- My mother and siblings for their continued support and encouragement, especially Mother for financial support during my, umm, 8 years as a student, and for looking after Anika, Oliver and Phrederyck.
- My friends, past and present, from the Departments of Genetics and Statistics; with particular thanks to Alison for putting up with me invading her office, sharing the ups and downs of th*s*s-itis, and for her sterling work proof-reading the final draft.
- Howard for encouragement, help with the flowcharts, and for bravely typesetting many of the equations in Chapter 6.
- The inmates of 8TP, especially Matthew for patiently explaining the mysteries of computers, and for supplying the coffee.

... and $\text{\$thanks}^{\{4\}}$ to *wrecked_diver@aifh.ed.ac.uk* for answering my plea for help with $\text{\LaTeX}$ for putting up with me and my emails, for being Awkward, and for just being there.

I would also like to thank Mike Talbot for the opportunity to work for S.A.S.S. during the last months of writing-up. This work was funded by the Science and Engineering Research Council.

Declaration

I declare that this thesis has been composed by myself
and that the work described in it is my own.

Table of Contents

1. Introduction	1
1.1 An Introduction to Population Structure	1
1.2 Inference of Relationship	3
1.3 The Problem Approached in This Thesis	6
1.4 Identification of Individuals	8
1.5 The Structure of This Thesis	12
 2. The Likelihood Method	 14
2.1 Introduction	14
2.2 The Relationships to be Considered	16
2.3 Method	19
2.3.1 Notation	19
2.3.2 Calculations	20
2.3.3 Family Patterns	22
2.3.4 A Sample of F families	27
2.3.5 Multiple Loci	29
2.4 Likelihood Ratios	29
2.5 Ordering of Individuals and Families	30
2.6 Summary	32

3. First Rank Inferences	33
3.1 Introduction	33
3.2 Details for a Sample of Two Individuals	34
3.2.1 The Likelihood for Each Relationship	34
3.2.2 The Probability of Correct Inference	38
3.2.3 A Closer Look at Two Loci	48
3.2.4 Field Parameters for Two Individuals	54
3.3 A Sample of Three Individuals	57
3.3.1 The Likelihood for Each Relationship	57
3.3.2 The Probability of Correct Inference	61
3.3.3 Field Parameters for Three Individuals	70
3.4 More than Three Individuals	72
3.5 Summary	77
4. Further Inferences	80
4.1 Introduction	80
4.2 The Effect of Exclusions and Draws	81
4.2.1 Exclusion	81
4.2.2 Equal Ranking	87
4.3 The Likelihoods and Likelihood Ratios for Valid Relationships	89
4.3.1 Expected Likelihood Ratios	90
4.3.2 Observed Likelihoods and Likelihood Ratios	94
4.3.3 Likelihood Ratios for Relationships with Several Families	98
4.4 Patterns of Inference	99
4.5 Summary	107

5. The Method Applied to Single Samples	110
5.1 Introduction	110
5.2 Analysis of Single Simulated Samples	111
5.3 An Example for a Natural Population	123
5.4 Summary	132
6. Analysis of Variance Method	134
6.1 Introduction	134
6.2 The Model	136
6.3 Analysis of Variance Based on Allele Scores	137
6.3.1 The Expected Mean Square Between Rots	138
6.3.2 The Expected Mean Square Within Rots	141
6.3.3 Results from Simulations Experiments	144
6.4 Analysis of Variance Based on Genotype Scores	150
6.4.1 The Expected Mean Square Between Rots	150
6.4.2 The Expected Mean Square Within Rots	152
6.4.3 Results from Simulation Experiments	154
6.5 Analysis of Variance Based on Number of Heterozygotes	157
6.5.1 The Expected Mean Square Between Rots	160
6.5.2 The Expected Mean Square Within Rots	161
6.5.3 Results from Simulation Experiments	161
6.6 Combining Results	162
6.7 Summary	167
6.8 Illustrative Example of Family Size Estimation Applied to Real Data . .	168
6.8.1 Summary of the Data	169
6.8.2 Family Size Estimation	169

7. Conclusions **173**

 7.1 Summary 173

 7.1.1 The Likelihood Method 174

 7.1.2 The Analysis of Variance Method 178

 7.2 Conclusions and Extensions 179

A. The Likelihood Equations **183**

B. Computer Programs **195**

 B.1 The Likelihood Method Programs 197

 B.2 Analysis of Variance Program 207

Bibliography **208**

Chapter 1

Introduction

1.1 An Introduction to Population Structure

Interpretation of population structure is a major objective in many aspects of population biology, including studies of social organisation, mating behaviour, seed dispersal, migration distances and variation in reproductive success. The structure of a population is defined by the pattern of relationships between the individual units within it; these units may be subpopulations or single organisms depending on the level of interest.

The *detailed* structure of a population can be inferred by examining the pattern of genealogical relationships between individual organisms. Most investigations into population biology therefore require an accurate assessment of these relationships as the foundation for further research. Given enough data on the genotypes of individuals it should be possible to infer the genetic relationships between them and thereby reconstruct their recent genealogical history. This knowledge can then be used to draw conclusions about the individuals and perhaps the population as a whole.

Methods for the inference of relationship have been most actively pursued in the field of human genetics and medicine. Genotypic or phenotypic data can be used to reconstruct pedigrees, and the genotypes of missing individuals can be inferred; the genotype of founder members can be inferred retrospectively, and predictive studies can be carried out in which the phenotype of an unborn individual is deduced from known relatives. Often, an accurate assessment of pedigree structure is required just among

living individuals; anecdotal evidence of family trees in small-scale human societies has been found to be only 80% reliable [Thompson, 1975]. Identification or confirmation of pedigree is also often necessary in domestic species, in progeny testing for example, or to establish that an individual belongs to a particular family. It is also useful to be able to find or confirm the relatedness of individuals within a breeding program either to avoid or promote consanguineous matings (relatedness is used as a general term for the level of genetic similarity between individuals, based on their sharing alleles identical by descent).

Turning to the field of forensic science, there are two fundamental questions that may be asked regarding genetic material: *relationship* (a major application is the exclusion of paternity) and *identity* (does a genetic sample come from individual 'A'?). The question of identity may be viewed as an extreme form of relationship inference between an individual and a sample of genetic material. Exclusion as well as confirmation of relationship can be useful; current techniques, such as DNA fingerprinting, claim to be able to establish paternity beyond reasonable doubt but it is often sufficient for a jury to know that a given suspect cannot be excluded from suspicion.

More generally, the fitness of plants or animals within a population is largely determined by the number of offspring they produce and these offspring must therefore be identified and counted. Differential fitness is a fundamental component in evolution. In plant populations, for example, if the spatial locations of the parent plants can be deduced in relation to the location of their seedlings, then this information on parentage can be used to study pollen or seed dispersal patterns. Life-history studies and population dynamics are the primary concern from an ecological perspective, and from a genetic perspective the mating system and its relationship to allele frequencies are of interest. Measurement of relatedness between interacting individuals is also a vital part of testing sociobiological theories.

Although information on population structure is utilised in many areas of research, the basic approach usually requires the identification of individuals by genotype or phenotype, followed by the estimation of the relationship between them based on this information.

1.2 Inference of Relationship

The methods used to identify the genotype of an individual are described in Section 1.4. In this section, previous research is outlined to illustrate some of the ways in which relatedness and population structure have been investigated and to give examples of how the results can be interpreted for a range of species. The limitation of these methods that motivated the work presented in this thesis is introduced.

Reconstruction of genealogies has classically involved human populations and has primarily been used to test genetic hypotheses regarding mode of inheritance and for genetic counselling. The techniques can equally be applied to other diploid species, both plants and animals. Pedigree data usually consists of data collected from one or more groups of related individuals extending over several generations. If the relationships are known, then the pedigree can be examined to establish the presence or absence of a genetic mechanism for a particular trait, i.e. to establish if there is a simple dominant or recessive mode of inheritance [Elston & Stewart, 1971]. Conversely, if the pattern of inheritance is assumed to be known then this information can be used, together with observed genotypes, to make inferences about the pedigree. Risk analysis is concerned with the derivation of probabilities that unborn individuals will have certain genotypes (or phenotypes) conditional on the phenotypes of known relatives. Similar probabilities can be found for ancestral inference; the genotypes of ancestors (the founders of the pedigree) are of interest and can be inferred from the genotypes of living descendants [Thompson *et al*, 1978].

When reconstructing genealogies, the first step is often to find possible offspring-parent links or offspring-parent-pair triplets. Consider a set of individuals whose genotypes are known for several loci. Taking each individual in turn, it is possible to evaluate the statistical likelihood of other individuals being either the maternal or the paternal parent. Alternatively, male-female pairs can be considered jointly as the parent-pair of the proposed offspring; each individual is grouped with one or two others that are most compatible as parents on the basis of sex, age and genotype. Thus, entire pedigrees can be reconstructed by linking together the pairs or triplets sequentially. Thompson [1976] also approached reconstruction of human pedigrees by

identifying pairs of parents and full sibships in addition to simple triplets of parents and one offspring. Meagher and Thompson [1986] developed strategies of genealogical reconstruction using single parent-offspring likelihoods. They found that if the true parents of a proposed offspring were present in the data set, then the pedigree could be accurately reconstructed. The method used by these authors and the results obtained are discussed further in Section 2.1. For natural populations, however, the number of possible parents may be large making exhaustive evaluation of all triplets a time consuming or impossible process.

Confirmation of pedigree using observed genotype can be useful for laboratory populations. Squirrel monkeys (genus *Saimiri*) are frequently used as models for biomedical and behavioural research and accurate pedigree details are often required. Although laboratory records are kept, errors occur particularly as squirrel monkeys are unusual in that allomaternal behaviour is quite common; the rearing of infants is shared among females so the biological mother cannot be assumed from observation [VandeBerg *et al*, 1990]. By studying a set of 14 polymorphic loci in 89 progeny, VandeBerg *et al* identified possible pedigree record errors, and in some cases were able to infer the likely correct pedigree. In each case one or both parents of the 89 monkeys were present.

While extensive pedigrees can be reconstructed by successive acceptance of pairs or triplets, often it is only necessary to consider two generations, for example in examining the mating behaviour in wild populations. Many bird species are thought to be monogamous, forming stable pairs to raise offspring. Starch gel electrophoresis of seven polymorphic loci carried out in a population of house sparrows revealed that 8% of nestlings cannot genetically be the offspring of both nest attendants [Wetton *et al*, 1987]. This type of evidence was used with behavioural observation to detect egg-dumping (intraspecific brood parasitism) or infidelity, by Quinn *et al* [1987] using restriction fragment length polymorphisms (RFLPs) in lesser snow geese (see also Chapter 5). In a similar investigation for a population of dunnocks, DNA fingerprints were used to link observation of mating behaviour and subsequent parental care with precise measurements of reproductive success. In this example, the species were known to be polyandrous but more detailed behavioural studies were made possible by accurate identification of offspring [Burke *et al*, 1989].

There have been many investigations into the establishment of paternity. In an extensive study based on 1,702 disputed paternity cases, Jeffreys *et al* [1991] examined the properties of human DNA fingerprints as resolved by two multilocus probes (named 33.6 and 33.15). In each case, the data consisted of a trio of mother, child and alleged father, data from other children being discarded. When the mother was absent, it was possible to consider the proportion of bands (see Section 1.4) for the child that were shared by the alleged father. Case-studies showed that 14% of bands were shared by unrelated individuals, whereas 57% were shared by a parent and child. So by comparing the male's genotype with the child's, paternity could be established. Another area in which it is necessary to assess accurate genetic relatedness is that of multiple paternity. A study carried out within a population of ground squirrels showed that 78% of litters were multiple sired. This was achieved by examining proteins from six loci and comparing the mother and offspring with potential fathers [Hanken & Sherman, 1981]. Foltz and Hoogland [1981] calculated the likelihood of paternity for their analysis of the mating system of the prairie dog. This was based on the probability of obtaining the data on parent-offspring trios given that the relationship was the correct one. Multiple insemination may also provide an opportunity for sperm competition when the female can store sperm [Williams & Evarts, 1989]. For example, commercially reared turkeys are artificially inseminated with sperm from as many as twelve males. It is of interest to examine the offspring and estimate the number of males that contributed genetic material. Assessment of numbers of mates and offspring allows evaluation of reproductive success, and this in turn can be used to identify correlations between reproductive success and phenotype or genotype.

In plant populations, if the spatial locations of the parent plants are known in relation to the location of the seedlings, then this information on parentage can be used to study pollen dispersal patterns. Meagher and Thompson [1987] located sets of possible parents on the basis of ecological criteria; genetic likelihoods of parent and parent-pair combinations were then examined using likelihood models. These models were based on the joint genotype probabilities of a specified offspring-single parent pair or offspring-parent pair triplets. A log-likelihood ratio (LOD score) of the joint genotype probabilities given parentage or unrelated was obtained for each pair or triplet in the set, and the parent or parent-pair having the highest LOD score was called the

maximum likelihood parent or parent pair. Here ecological and genetic data were both incorporated into the process of genealogical reconstruction, with the goal of identifying the most likely maternal and paternal parents for seedlings undergoing natural establishment in a population of *Chamaelirium luteum*. This type of spatial distribution information can be incorporated into tree improvement programs, for example [Schwarzmann & Gerhold, 1991].

The unifying link between the above examples is the availability of data from one or both parents. This, however, is only possible when adequate data on members of the parental generation are available for analysis. In natural populations, such as *Drosophila* or barnacles, the parents will not be present when the offspring emerge. In other species, full sibships may not be easily identifiable by observation of behaviour and cannot be assumed *a priori*. For birds and mammals, for example, half siblings may sometimes be assumed by the presence of a female parent, but the male parent of each offspring could be unknown. This is particularly a problem when studying ephemeral species such as insects, non-social species where parents are not seen to attend their young, or in plants where offspring may be widely dispersed.

1.3 The Problem Approached in This Thesis

A full analysis of a many-generation pedigree requires a considerable amount of data to be collected from many individuals. In short-lived or dispersed species this may be at best difficult or even impossible. Many studies have therefore been carried out using families extending over just two or perhaps three generations.

Direct comparison of genotypes between two or more generations is not always feasible however. In natural populations lacking familial structure, one or possibly both parents are unknown. The question then arises whether useful information regarding genetic relationships can be obtained from analysis of data gathered from members of a single generation when there is no information regarding the immediate ancestors. The analysis presented here is an estimation of the relationships between individuals on the basis of their genotypes at a number of loci, for the particular case where

all individuals are from the same generation. From this starting point more general inferences on population structure can be made.

This situation is applicable to a wide range of experiments both in the laboratory and in the field; precise interpretation of the results will depend on the objectives of the research being carried out. In this thesis, therefore, two general situations are envisaged as follows:

- **Identification of Individual Relationships**

For human families, birds or small mammals, the identity of each individual in the sample may be important. The experimenter may ask, "Is **A** the sibling of **B** or **C**?". Experiments of this class will consist of small numbers of individuals (less than ten) where the number of possible combinations of relationships is small.

The "likelihood approach" investigates the likelihood for different relationships among individual members of a sample from the population. It is assumed that the detailed relationship between these individuals is important and that a small number of named individuals are considered simultaneously. This is based on the method developed by Thompson [1976] and uses the probability of obtaining the genotypes observed in the sample, given that the hypothetical relationship under consideration is the true one.

- **Estimation of Family Size**

An alternative situation arises if the sample size is large, or where details of exact genealogical relationship are not required. In human populations, emphasis would be placed on identification of siblings as individuals; in other populations it may be sufficient to be able to subdivide the group into smaller units each consisting of individuals that are more closely related to each other than to members of other groups. This would enable some estimation to be made regarding the numbers of parents contributing to the next generation. Experiments using plant material might be included in this class, or perhaps field experiments with insects. For example, how many families are represented in a sample of flies emerging from a single banana? Here, the investigation is essentially an estimation of the mean relationship in the sample of individuals.

1.4 Identification of Individuals

Several methods have been used to identify individuals and their relationships with others; examples include analysis of many tissue and blood types, and morphological features such as mandibular shape [Morton *et al*, 1987]. Generally, individual specificity is not essential for population studies, but it is necessary to be able to estimate the probability that a random individual has a particular genotype or phenotype.

Most work has been carried out using polymorphic enzyme loci. Biochemical and molecular techniques, such as enzyme electrophoresis, have been extended, improved and successfully applied to many areas of evolution and population biology. Allozymes and blood proteins tend to have a clear Mendelian inheritance pattern, but generally show limited variability so that many individuals have identical genotypes, thus restricting their usefulness as markers. Recent advances in molecular techniques have, however, made it possible to examine DNA sequences directly and a wealth of information has been revealed. In addition, the development of such techniques as the polymerase chain reaction (PCR) has enabled DNA samples to be amplified so that many analyses for different markers can be carried out starting with a very small (sub-nanogram) amount of genomic DNA [Weber & May, 1989].

Analysis of an individual's DNA begins by cleaving the genomic DNA with restriction enzymes. Fragments of DNA are created that vary in length depending on the number and position of the appropriate restriction sites; RFLPs are revealed after hybridisation to labelled probe and separation by gel electrophoresis. The resultant banding pattern on the gel (or autoradiograph) reveals the DNA profile of the individual; each band represents DNA fragments of similar length.

While the essential coding regions of the genome cannot vary greatly between individuals, non-coding regions have revealed extensive polymorphism. At some loci, short DNA sequences are found to occur in tandem repeats; a core sequence of approximately 9–40 base pairs in length (usually GC rich) can be repeated thousands of times [Jeffreys *et al*, 1985b; Debenham, 1992]. These are termed minisatellites and are hypervariable; unrelated individuals differ greatly in the number of repeat units at each of these VNTR (variable number of tandem repeat) loci. If the restriction enzymes are chosen so that

they do not cut the core sequence, then the number of repeat sequences determines the overall length of the DNA fragments; each different length defines an allele of the VNTR locus. The alleles are inherited in a simple codominant Mendelian manner. Each allele will ideally be visible as a distinct band on the electrophoretic gel, but in practise gel resolution is such that the distribution of fragment lengths is essentially continuous, not discrete, and classification into alleles is not straightforward [Budowle *et al*, 1991]. However, statistical methods are available to estimate allele frequencies allowing for both measurement error and coalescence between bands of similar size [Devlin *et al*, 1991]. Also, the lanes of a gel can be divided into small zones, called bins, such that bands within this region are said to represent the same allele. The fixed-bin method of counting alleles has been shown to be reliable [Weir, 1992].

A hypervariable locus can be used as a probe to detect other loci with a similar sequence. Under high stringency conditions, single locus probes (SLPs) are used to detect a particular genetic locus; the probe binds to only one locus and thus reveals only one band (for homozygotes) or two bands (for heterozygotes) for each individual. DNA variability is examined at a defined genetic locus. Locus-specific minisatellite probes can be isolated efficiently from humans and many other species [Wong *et al*, 1987; Burke *et al*, 1991], and are particularly useful for linkage analysis. SLPs can be used singly or in cocktails to produce a DNA profile for an individual; generally, four or five separate SLPs are used in a forensic analysis or for paternity identification [Jeffreys *et al*, 1991b]. The choice of restriction enzyme and probe is important for detecting different levels of variability.

Under low stringency conditions, a hypervariable locus can be used as a generic probe to detect other loci that share a similar sequence (within or between species). These multilocus probes (MLPs) are used to detect many hypervariable loci simultaneously. Perhaps the most interesting and potentially useful discovery using VNTR loci is the development of DNA "fingerprints". Jeffreys *et al* [1985] developed DNA probes that hybridize to hypervariable minisatellite regions of genomic DNA to produce a banding pattern on a gel known as a DNA fingerprint, so named as these banding patterns are highly individual specific. Although originally developed in humans, the technique has proved useful in many other species, especially wild bird populations, but also domestic and laboratory animals [Hill, 1987; Jeffreys & Morton, 1987; Jeffreys

et al, 1991b]. The drawback with the use of DNA fingerprints for relatedness studies is that it is generally not known which bands belong to which loci; a fingerprint is thus a phenotype not a genotype. In addition, the exact number of bands scored on a gel is determined by the resolution of the gel, and with the technology currently available errors in measurement occur. Fragments of DNA with the same mobility, which are therefore recorded as the same band, are not necessarily isoallelic; some loci will produce more than one band on the gel and two or more loci may produce bands that cosegregate (leading to pseudohomozygosity); very small fragments may be lost from a gel and not scored (blank alleles). Thus the term *band* is not synonymous with *allele*. Many alleles may exist for a given locus, and if two markers from a single parent are found to be inherited by one offspring then allelism can be ruled out. However, there are generally so many bands segregating on a gel (and not all will be distinguishable) that allelism cannot be established. The proportion of shared marker bands between two or more individuals will not necessarily be equivalent to the proportion of shared genes, making accurate analysis very complex.

DNA fingerprints can be used to identify individuals uniquely and to establish parentage. Scene of crime samples can be analysed and used to exclude or implicate a suspect. Also, every band present in the fingerprint of an offspring occurs in at least one parent (barring mutation); thus a pair of individuals can sometimes be *excluded* as the parents of a child by comparing the offspring fingerprint with those of both putative parents [Gill *et al*, 1985; Jeffreys *et al*, 1985a; Wells *et al*, 1988; Wetton *et al*, 1987]. Paternity disputes are most simply investigated by counting the number of bands in the child that are not present in the mother, and finding how many of these unassigned bands can be detected in the alleged father. So while fingerprints can be used in simple cases by direct comparison of bands between two generations, inference of more distant relationships is difficult (see also Lynch [1988,1991] and Jeffreys *et al* [1991]). Further studies using DNA fingerprints rely on calculating the fraction of bands that are shared by two or more individuals. For example, there is known to be a linear relationship between the degree of inbreeding and band sharing in chickens [Kuhnlein *et al*, 1990]. Strains of chicken that are highly inbred will tend to share a large number of bands, whereas less inbred strains show greater heterozygosity and less band sharing.

Minisatellite alleles vary in repeat copy number, but may also vary internally due to variants of the basic unit being interspersed along the length of the locus. A technique known as MVR-PCR (minisatellite variant repeat) has been developed that allows the pattern of these variants, within a hypervariable locus, to be examined in human DNA [Jeffreys *et al*, 1991b; Jeffreys *et al*, 1991a]. MVR-PCR used on total genomic DNA produces a profile of two alleles superimposed, and a digital code is generated that pinpoints the differences between the alleles. Thus two alleles with the same length are distinguishable. It has been shown that the MVR-PCR codes are individual specific, and so this information is readily applicable to tests of both identity and paternity for example. This technique avoids many of the practical difficulties that limit the use of other fingerprinting techniques, and so promises to be a valuable source of genotype information.

Another potentially useful set of markers are microsatellites (or short tandem repeats); these are hypervariable loci whose repeat sequence is di-, tri- or tetra-meric. The number of repeat sequences varies between individuals so can be used in the same way as minisatellites although the number of repeats is much smaller. Microsatellites have been used successfully to confirm the identity of a victim in a forensic analysis using severely degraded DNA where minisatellite analysis was not possible [Hagelberg *et al*, 1991].

A novel DNA polymorphism assay was recently developed, based on the amplification by PCR of random DNA segments, using single (10-nucleotide) primers of an arbitrary nucleotide sequence [Williams *et al*, 1990; Welsh & McClelland, 1990]. The randomly amplified polymorphic DNAs (RAPDs) detect differences between individuals; these arise from changes in either the sequence of the primer binding site or from changes that alter the size or prevent amplification of a target DNA. The amplification products are transmitted as dominant markers, in contrast to RFLPs which are codominant. The attraction of RAPD analysis over the use of minisatellites is that it is technically simple, quick, requires small amounts of DNA, and does not need radioactively labelled probes. RAPDs are now being used in a wide range of species and are proving to be very useful for detecting novel polymorphism in species as diverse as wheat [Vierling & Nguyen, 1992], honey bees [Hunt & Page, 1992], potatoes [Baird *et al*, 1992] and fungi [Pellissier Scott *et al*, 1992].

For the analyses presented in this thesis, the data on which inferences are to be based are the genotypes of individuals for a number of unlinked autosomal loci. The sampled individuals are assumed to have been assayed for several marker loci, so that there is a genotype for each one, for example data from allozymes, single locus RFLPs or single-locus minisatellite probes. Throughout this thesis it is also assumed that each allele can be reliably distinguished so that all possible genotypes may be isolated and recorded if present in the population. Linkage equilibrium is also assumed. Simulation of appropriate data is carried out by means of a set of computer programs that are specifically written for this purpose. A single locus with two alleles provides the minimum level of information; loci are therefore simulated to have between two and twenty alleles. The accuracy of the inferences made regarding genealogy are compared using different numbers and types of loci. In addition to simple simulated data, inferences are made using a selection of "field parameters"; details of these loci have been taken from published papers and are chosen to represent the level of information typically available.

1.5 The Structure of This Thesis

This thesis comprises two separate but related parts. An introduction relevant to each is given at the beginning of Chapters 2 and 6.

The Likelihood Method

In Chapter 2, the type of relationships to be considered in the analysis are described. Details of the calculation of the likelihood for each hypothetical relationship given the data are also given, in general terms. The procedure for finding likelihood ratios and ranking the relationship hypotheses is described.

In Chapter 3, inferences of relatedness are made based solely on the relationship ranked first. Full details are given for a sample of two individuals, and the "probability of correct inference" is found for small sample sizes. Typical results are shown as obtained from computer simulation.

Inferences on relationship using the ranked likelihood ratios are again examined in Chapter 4, but here the range of the likelihood ratios is taken into account. Trends within the top-ranked structures are also identified. The effect of exclusion is discussed, and simulation results are presented for a range of parameters.

Finally for this part, in Chapter 5 some examples are given showing the interpretation of single sample data using the results obtained in the previous chapters. Examples are given for both simulated and field data.

The Analysis of Variance Method

In Chapter 6 the model population from which the sample individuals are to be drawn is described and an analysis of variance is carried out utilising three types of data: allele counts, genotype counts and scores based on heterozygosity. Typical results are shown using simulated sample data for each case. Section 6.8 describes the results obtained when this method of analysis is applied to a set of field genetic data and compares the conclusions reached here with those published.

Conclusions

Chapter 7 summarises and compares the achievements made using each of the techniques and considers their shortcomings. Possible improvements and extensions are discussed.

Chapter 2

The Likelihood Method

2.1 Introduction

Some problems associated with the inference of genetic relationship are introduced in Section 1.3. If members of the parental generation are not available, then alternative methods are required to deduce the relatedness of individuals in the sample.

This problem has been addressed, in part, by other authors. Williams and Evarts [1989] developed a technique to estimate the probability of concurrent multiple insemination. Their aim was to relate the probability of observing particular offspring genotypes (in terms of the number of copies of each of two alleles present) to two parameters: θ , the probability that a random offspring has a particular allele, and ϕ , the probability that a random offspring comes from a family that was multiply fathered. This was carried out both for the case where the mother's genotype was known and where it was unknown, i.e. both parents were unknown. However, their model used a two-allele collapse of the available data and assumed only single or double insemination. The individuals in the sample could therefore be assumed to be at least half siblings. Roeder *et al* [1989] estimated fertilities for parent-pairs, including the case where neither parent was known. Their primary interest, however, was in the proportion of offspring assigned to a given adult rather than a detailed pedigree.

For the first analysis presented in this thesis (named the likelihood method), it is assumed that a small number of diploid individuals are sampled from the same generation, fledglings from a particular nest perhaps, or insects emerging together in a specific

location. The data on which inferences of relationship are to be based are the genotypes of these individuals for a number of independent autosomal loci. An estimation of the detailed genealogical relationship between these individuals is required.

Relationships between individuals can be specified by the alleles they have in common; these alleles may be identical by descent (IBD) from some common ancestor. At any single autosomal locus, two diploid individuals may have zero, one or two alleles in common and these may be identical in state or IBD. There are nine distinct states of gene identity depending on the origin of the alleles. The probabilities of these nine states can be computed for any specific genealogical relationship between two individuals [Jacquard, 1972; Elandt-Johnson, 1971b; Thompson, 1980]. Thompson [1974] also extended this specification of relationship and derivation of joint genotype distribution to other relatives, including sibs, uncles and grandparents.

The probability of any joint set of genotypes can be computed for more than two individuals, given a particular genealogy. The probabilities for an observed set of genotypes under different genealogical hypotheses are defined as the likelihoods for these hypotheses; the genealogy of maximum likelihood may be found by comparing the likelihoods for different hypotheses. Thompson [1975] further considered the problem of inferring the relationship between two individuals (who may or may not be from the same generation), based on the likelihood for the relationship. Thompson defines the relationship between two individuals A and B as $x(A, B) = (\kappa_0, \kappa_1, \kappa_2)$, where κ_i is the probability of A and B sharing i genes identical by descent. Thus if A and B are offspring and parent, $x(A, B) = (0, 1, 0)$. If A and B have genotypes G_1 and G_2 respectively, then the likelihood (L) for the relationship between A and B is

$$\begin{aligned} L(x) &= \Pr(G(A) = G_1, G(B) = G_2 \mid x(A, B) = x) \\ &= \sum_{i=0}^2 \kappa_i \Pr(G(A) = G_1, G(B) = G_2 \mid i). \end{aligned}$$

This general approach to relationship inference is extended in this thesis for the relationship between N individuals from the same generation. The basic procedure is to determine a set of possible genetic relationships that could exist between members of a group, calculate the likelihood for each hypothetical relationship given the data, and then use these likelihoods to infer the genealogical structure in the sample.

2.2 The Relationships to be Considered

The first step in the analysis is to find the set of relationship structures that could exist within the group. Once these hypothetical relationships are established, the likelihood for each can be calculated. Within the sample, the types of relationship structures envisaged depend on the number of parent-pairs that contribute to the offspring generation. It is assumed that each parent mates at random with only one other individual, so that the offspring population consists exclusively of groups of either full siblings or unrelated individuals. Hence, for any sample size N there is a defined series of relationships; these relationships, and the notation used to distinguish them, are described in this section. The problem is oversimplified by considering only these relationship structures. In most real situations other relatives will be included, the most common being half siblings due to polygyny and/or polyandry. This method could, however, be extended to include these alternative relationships by distinguishing between male and female parents in the probability calculations.

When strict monogamy is assumed to be the pattern of mating in the parental generation, there are a limited number of ways in which the members of the sample can be related to each other. A family is defined to be a group of full siblings; a sample of size N may represent F such families, with $F = 1, 2, \dots, N$ and n_i members in family i where $\sum_{i=1}^F n_i = N$ and $n_i \neq 0$.

For example, suppose a sample consists of four individuals, labelled X_1 to X_4 . The relationships considered for four individuals in this model are as follows:

- | | | |
|---|----|---------------------------|
| one family with four full siblings; | 1) | $X_1 X_2 X_3 X_4$, |
| two families, with three in one family; | 2) | $X_1 X_2 X_3 \quad X_4$ |
| | 3) | $X_1 X_2 X_4 \quad X_3$ |
| | 4) | $X_1 X_3 X_4 \quad X_2$ |
| | 5) | $X_2 X_3 X_4 \quad X_1$, |
| or two families each with two siblings; | 6) | $X_1 X_2 \quad X_3 X_4$ |
| | 7) | $X_1 X_3 \quad X_2 X_4$ |
| | 8) | $X_1 X_4 \quad X_2 X_3$, |

or three families;

- | | | | |
|-----|----------|-------|---------|
| 9) | X_1X_2 | X_3 | X_4 |
| 10) | X_1X_3 | X_2 | X_4 |
| 11) | X_1X_4 | X_2 | X_3 |
| 12) | X_2X_3 | X_1 | X_4 |
| 13) | X_2X_4 | X_1 | X_3 |
| 14) | X_3X_4 | X_1 | X_2 , |

or four families each with one member; 15) X_1 X_2 X_3 X_4 ,

making a total of 15 possible relationships. Note that to describe the *relationship*, the order of individuals within families and the order of families is not important.

As N increases, the number of potential relationships increases rapidly; there are 5 such structures for a sample of size 3, and 15, 52, 203, 4140, 115975 and 4213597 for samples of size 4, 5, 6, 8, 10 and 12 respectively (see Figure 2.1). These numbers are found by combining the appropriate *Stirling numbers of the second kind* that are tabulated in many mathematical texts e.g. Abramowitz & Stegun [1965] as $S_N^{(F)}$, where F is the number of families and N the sample size here. For $N = 4$ as above, $S_4^{(1)} = 1$, $S_4^{(2)} = 7$, $S_4^{(3)} = 6$ and $S_4^{(4)} = 1$, making a total of 15.

While $S_N^{(F)}$ gives the total number of ways of partitioning N individuals into F (non-empty) families, clearly these groups can be further subdivided according to the size of the families. For large sample sizes it is impractical to list and count all possible relationships. The number of relationships included in each of the groups is equal to the number of ways of placing N different individuals into F families with n_i in the i^{th} family (without regard to the order of families). This is equal to

$$\frac{N!}{n_1!n_2!\dots n_F!} \frac{1}{k_1!k_2!\dots},$$

where k_1, k_2 , etc represent the number of families indistinguishable with respect to the number of siblings. For example, for $N = 6$ and $F = 4$ the sample may be four families of size 3, 1, 1 and 1, or of size 2, 2, 1 and 1. Hence,

$$S_6^{(4)} = \left(\frac{6!}{3!1!1!1!} \cdot \frac{1}{3!} \right) + \left(\frac{6!}{2!2!1!1!} \cdot \frac{1}{2!2!} \right) = 20 + 45 = 65.$$

So, for a given sample of N individuals, the total number of relationships to be considered is

$$\sum_{F=1}^N S_N^{(F)}, \text{ where } S_N^{(F)} = \sum_{\Omega(F,N)} \left(\frac{N!}{n_1!n_2!\dots n_F!} \cdot \frac{1}{k_1!k_2!\dots} \right) \quad (2.1)$$

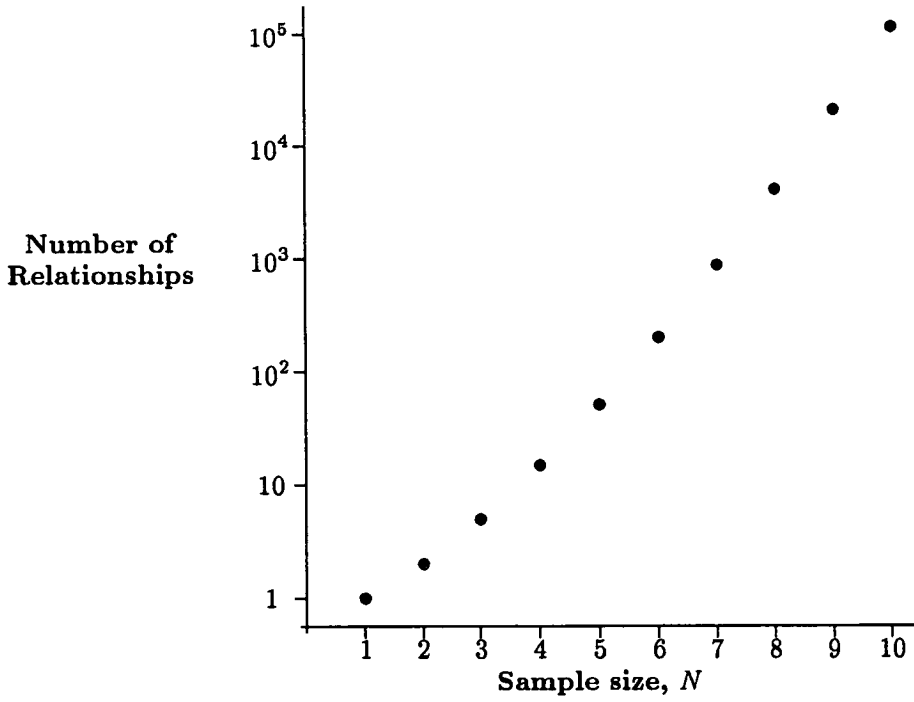


Figure 2.1: The total number of potential relationships increases rapidly with increasing sample size, N . (note 1 to 10^5 scale)

and where $\Omega(F, N)$ is the set of F integers n_i such that $\sum n_i = N$, $n_i \neq 0$.

Throughout this analysis, the following notation is used to describe the relationship structure under consideration. While individuals are labelled X_i , groups of full siblings are enclosed by braces $\{ \}$ to indicate families. For example,

$$\{X_1, X_3, X_6\}\{X_2, X_5\}\{X_4\}$$

indicates a sample of size six with three families in which individuals numbered 1, 3 and 6 are full siblings, individuals numbered 2 and 5 are full siblings from a second family and number 4 is the single member of family three. Relationships are also denoted $R_{F:n_1, n_2, \dots, n_F}$, so the example given above corresponds to $R_{3:3, 2, 1}$.

Following this notation, a ‘wild-card’ ‘*’ is used to denote similar relationships. For example, $R_{2:*}$ refers to all possible arrangements that consider the individuals as belonging to two families.

For this simple model, sampled individuals may either be full siblings or not closely related to each other. In other words, “unrelated” individuals are those that have no

parents in common. In a natural population, other relationships are of course possible, half siblings being the most common. This is discussed further in Chapter 7.

2.3 Method

Having found the possible relationships that might exist within the group, alternative hypotheses are compared by calculating the likelihood for each relationship. The likelihood for a particular relationship is calculated as the probability of obtaining the observed genotype data given that the relationship under consideration is the true one [Thompson, 1975]. This probability clearly depends on both the genotype array in the sample and on the conditional probability of its occurrence with given parental mating types.

2.3.1 Notation

The sample consists of N diploid individuals, drawn from a large panmictic population. Genotypes are known for the individuals in the sample for m unlinked autosomal loci. For this analysis it is assumed that the sample contains members of one generation only, and that all individuals are non-inbred and in Hardy-Weinberg equilibrium. The genotype of individual i at locus j is given by γ_{ij} , whilst $\Gamma_i = [\gamma_{i1}, \dots, \gamma_{im}]$ is the genotype of individual i over the m loci. The number of alleles segregating at locus j is s_j for $j = 1, 2, \dots, m$. It is assumed that the frequency of each allele in the population is known; the frequency of allele $A_{j,i}$ is $p_{j,i}$ for $i = 1, 2, \dots, s_j$ with $\sum_{i=1}^{s_j} p_{j,i} = 1$ for each locus.

As described in Section 2.2, members of the sample are labelled $X_1, X_2, \dots, X_N$ and a family is defined as a group of individuals that have both parents in common. The sample to be analysed can be considered as belonging to F such families ($F = 1, 2, \dots, N$) with n_i members in family i , where $\sum_{i=1}^F n_i = N$, ($n_i \neq 0$). The notation $R_{F:n_1, n_2, \dots, n_F}$ is used to indicate a hypothetical relationship structure with F families.

Parent Pair ($\gamma_u \times \gamma_v$)		$\alpha_{u,v}$
1.	$A_i A_i \times A_i A_i$ Both parents identical homozygotes	p_i^4
2.	$A_i A_i \times A_j A_j$ Different homozygotes	$2p_i^2 p_j^2$
3.	$A_i A_i \times A_i A_j$ One homozygote, one heterozygote with one allele in common	$4p_i^3 p_j$
4.	$A_i A_i \times A_j A_k$ One homozygote, one heterozygote with no alleles in common	$4p_i^2 p_j p_k$
5.	$A_i A_j \times A_i A_j$ Both parents identical heterozygotes	$4p_i^2 p_j^2$
6.	$A_i A_j \times A_i A_k$ Two heterozygotes with one common allele	$8p_i^2 p_j p_k$
7.	$A_i A_j \times A_k A_l$ Two heterozygotes with no common alleles	$8p_i p_j p_k p_l$

Table 2.1: The seven general parental mating types and their unordered frequencies, $\alpha_{u,v}$, assuming random mating. For each type $i, j, k, l = 1, 2, \dots, s, i \neq j \neq k \neq l$.

2.3.2 Calculations

This section gives details of the calculation of the likelihood for each hypothetical relationship for the sample of N individuals.

Consider a family consisting of n siblings. The genotype of each individual in this family is determined by the genotypes of the parents; in this analysis we know the genotypes of the sampled offspring but not those of the parents.

For any given individual, maternally and paternally inherited alleles cannot normally be distinguished and since only autosomal loci are considered, no distinction is made between the two parents here. The allele frequencies are assumed to be equal in males and females. The unknown parental genotypes are denoted Γ_u and Γ_v , where $\Gamma_u = [\gamma_{u1} \gamma_{u2} \dots \gamma_{um}]$. For a single locus with s alleles, seven types of parent-pair matings can be distinguished, based on the number of alleles the male and female have in common. These are shown in Table 2.1, together with the relative frequency, $\alpha_{u,v}$, with which each pair occurs under the assumption of random mating. Each of these pairs ($\gamma_u \times \gamma_v$) can produce offspring of known genotype; the conditional probabilities of offspring given these parental matings are shown in Table 2.2. For example, the

Parent Pair	Offspring Genotypes						
	$A_i A_i$	$A_j A_j$	$A_i A_j$	$A_i A_k$	$A_j A_k$	$A_i A_l$	$A_j A_l$
1. $A_i A_i \times A_i A_i$	1	0	0	0	0	0	0
2. $A_i A_i \times A_j A_j$	0	0	1	0	0	0	0
3. $A_i A_i \times A_i A_j$	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0	0	0
4. $A_i A_i \times A_j A_k$	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
5. $A_i A_j \times A_i A_j$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0	0	0	0
6. $A_i A_j \times A_i A_k$	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0
7. $A_i A_j \times A_k A_l$	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Table 2.2: The conditional probabilities of offspring given parental mating type.

probability that an individual has genotype $A_i A_j$ conditional on the parental genotypes $A_i A_j$ and $A_i A_k$ is $1/4$. As the genotypes of successive offspring are independent the probability of obtaining any *ordered* set of n offspring $(\gamma_1 \gamma_2 \dots \gamma_n)$ given parents $\gamma_u \times \gamma_v$ is

$$\prod_{i=1}^n \Pr(\gamma_i \mid \gamma_u \times \gamma_v), \quad (2.2)$$

and the probability of obtaining a family of n siblings with these genotypes in a population is the sum over all possible parent pairs. Consider a locus with s alleles; there can be s homozygotes $A_i A_i$ and $\frac{1}{2}s(s-1)$ heterozygotes $A_i A_j$, making a total of $\frac{1}{2}s(s+1)$ genotypes. If these genotypes are numbered 1 to g , then for n siblings

$$\begin{aligned} \Pr(\gamma_1, \gamma_2, \dots, \gamma_n \mid R_{1:n}) &= \sum_{\gamma_u=1}^g \sum_{\substack{\gamma_v=1 \\ \gamma_u \leq \gamma_v}}^g \Pr(\gamma_1, \gamma_2, \dots, \gamma_n \text{ and } \gamma_u \times \gamma_v) \\ &= \sum_{\gamma_u=1}^g \sum_{\substack{\gamma_v=1 \\ \gamma_u \leq \gamma_v}}^g \alpha_{u,v} \prod_{i=1}^n \Pr(\gamma_i \mid \gamma_u \times \gamma_v). \end{aligned} \quad (2.3)$$

Note that the condition $\gamma_u \leq \gamma_v$ must be included as the frequency $\alpha_{u,v}$ takes account of different orders of the parents.

For example, if a locus has two alleles A_1 and A_2 and $\gamma_1 = A_1 A_1, \gamma_2 = A_1 A_2$ and $\gamma_3 = A_1 A_2$, the probability that a family of three siblings drawn from a population of

such families would have these ordered genotypes is

$$\begin{aligned}
 \Pr(\gamma_1, \gamma_2, \gamma_3 \mid R_{1:3}) &= \sum_{\gamma_u=1}^3 \sum_{\substack{\gamma_v=1 \\ \gamma_u \leq \gamma_v}}^3 \alpha_{u,v} \prod_{i=1}^3 \Pr(\gamma_i \mid \gamma_u \times \gamma_v) \\
 &= \sum_{\gamma_u=1}^3 \sum_{\substack{\gamma_v=1 \\ \gamma_u \leq \gamma_v}}^3 \alpha_{u,v} \Pr(A_1 A_1 \mid \gamma_u \times \gamma_v) [\Pr(A_1 A_2 \mid \gamma_u \times \gamma_v)]^2 \\
 &= 4p_1^3 p_2 \left(\frac{1}{2}\right) \left(\frac{1}{2}\right)^2 + 4p_1^2 p_2^2 \left(\frac{1}{4}\right) \left(\frac{1}{2}\right)^2 \\
 &= \frac{1}{4} p_1^2 p_2 (1 + p_1). \tag{2.4}
 \end{aligned}$$

Expression (2.4) shows the probability of obtaining the observed, ordered genotype data given the relationship $R_{1:3}$ and is called the likelihood for this relationship.

2.3.3 Family Patterns

Within any given family, for a single locus, there can be a maximum of four alleles because a set of full siblings must by definition have the same two parents, each with two copies of the gene. So assuming no mutation, only a subset of the g genotypes may be represented in a family for a given locus. There are in fact only 14 ‘family patterns’ (denoted Δ_i , $i = 1, 2, \dots, 14$), each pattern being defined by the number of different homozygotes and heterozygotes present and the number of alleles they have in common (see Table 2.3). For example, pattern Δ_1 has n individuals that are all identical homozygotes $A_i A_i$, whereas pattern Δ_{14} is characterised by r_1 members with genotype $A_i A_k$, r_2 with genotype $A_i A_l$, r_3 with genotype $A_j A_k$ and r_4 with genotype $A_j A_l$. For each pattern $\sum r = n$, $r \neq 0$ and $i \neq j \neq k \neq l$. A set of genotypes that do not fit one of these patterns cannot be members of a single family; it is not possible for three different homozygotes to be full siblings for example.

Within each pattern Δ_i , the individual genotypes may be ordered in a number of ways. As the genotypes of the siblings are independent, the probability associated with each ordered sample is the same. When calculating likelihoods for more complex relationships with more than one family, it is convenient to consider the individuals as ordered, therefore throughout the following chapters individuals or genotypes are considered as ordered unless otherwise stated (see also Section 2.5).

Pattern	Genotypes in Family						
	$A_i A_i$	$A_i A_j$	$A_i A_k$	$A_i A_l$	$A_j A_j$	$A_j A_k$	$A_j A_l$
Δ_1	n	—	—	—	—	—	—
Δ_2	—	n	—	—	—	—	—
Δ_3	r_1	—	—	—	r_2	—	—
Δ_4	r_1	r_2	—	—	—	—	—
Δ_5	r_1	r_2	—	—	r_3	—	—
Δ_6	r_1	—	—	—	—	r_2	—
Δ_7	—	r_1	r_2	—	—	—	—
Δ_8	r_1	r_2	r_3	—	—	—	—
Δ_9	r_1	r_2	—	—	—	r_3	—
Δ_{10}	r_1	r_2	r_3	—	—	r_4	—
Δ_{11}	—	r_1	r_2	—	—	r_3	—
Δ_{12}	—	—	r_1	r_2	—	r_3	—
Δ_{13}	—	—	r_1	—	—	—	r_2
Δ_{14}	—	—	r_1	r_2	—	r_3	r_4

Table 2.3: A family of n siblings can be classified as one of fourteen family patterns. Each pattern is characterised by the number of different homozygotes and heterozygotes ($i, j, k, l = 1, 2, \dots, s, i \neq j \neq k \neq l, \sum r_i = n$, for each set and $r_i \neq 0$).

The probability of obtaining an ordered sample of n individuals with genotype pattern Δ_i given that they are all full siblings can be found for each of these 14 types from Expression (2.3) for a single locus with s alleles. These expressions can be simplified and are the joint genotype distributions for n siblings.

For illustration, details are shown below for calculating the joint genotype distribution for a family fitting pattern Δ_1 . The other results are given briefly here, with full details given in Appendix (A).

Family Pattern Δ_1 The family pattern Δ_1 is defined as a family of n full siblings who are all identical homozygotes $A_i A_i$ for the locus of interest. The genotypes of possible parents and the conditional probability of obtaining these offspring are shown

in Tables 2.1 and 2.2. The relevant pairs that can produce $A_i A_i$ offspring are given below. The probabilities must be summed over all possible allele combinations. Note that for parent pair type number 6 when summing over all j and k , the probability must either be divided by two or $j < k$ must be specified, to avoid counting each parent pair twice; the frequency of parent pairs taken from Table 2.1 takes account of different orders of the parents.

Parental Types	$\alpha_{u,v}$	$\Pr(\Delta_1 \text{ and } \gamma_u \times \gamma_v)$
1. $A_i A_i \times A_i A_i$	p_i^4	$1^n p_i^4$
3. $A_i A_i \times A_i A_j$	$4p_i^3 p_j$	$\sum_{\substack{j=1 \\ j \neq i}}^s \left(\frac{1}{2}\right)^n 4p_i^3 p_j$
5. $A_i A_j \times A_i A_j$	$4p_i^2 p_j^2$	$\sum_{\substack{j=1 \\ j \neq i}}^s \left(\frac{1}{4}\right)^n 4p_i^2 p_j^2$
6. $A_i A_j \times A_i A_k$	$8p_i^2 p_j p_k$	$\frac{1}{2} \sum_{\substack{j=1 \\ k \neq j \neq i}}^s \sum_{k=1}^s \left(\frac{1}{4}\right)^n 8p_i^2 p_j p_k$

To find the overall probability for family pattern Δ_1 , the expressions must be summed over all possible parent pairs (Expression (2.3)),

$$\Pr(\Delta_1 \mid R_{1:n}) = \left(\frac{1}{4}\right)^{n-1} p_i^2 \left[4^{n-1} p_i^2 + 2^n p_i \sum_{\substack{j=1 \\ j \neq i}}^s p_j + \sum_{\substack{j=1 \\ j \neq i}}^s p_j^2 + \sum_{\substack{j=1 \\ k \neq j \neq i}}^s \sum_{k=1}^s p_j p_k \right].$$

So,

$$\sum_{\substack{j=1 \\ j \neq i}}^s p_j = 1 - p_i \text{ and } \sum_{j=1}^s \sum_{k=1}^s p_j p_k = \left(\sum_{\substack{j=1 \\ j \neq i}}^s p_j\right)^2 + \sum_{\substack{j=1 \\ j \neq i}}^s p_j^2 \text{ for } k \neq j \neq i,$$

and therefore,

$$\begin{aligned} \Pr(\Delta_1 \mid R_{1:n}) &= \left(\frac{1}{4}\right)^{n-1} p_i^2 \left[4^{n-1} p_i^2 + 2^n p_i (1 - p_i) + (1 - p_i)^2 \right] \\ &= \left(\frac{1}{4}\right)^{n-1} p_i^2 \left[p_i^2 (4^{n-1} - 2^n + 1) + 2p_i (2^{n-1} - 1) + 1 \right] \\ &= \left(\frac{1}{4}\right)^{n-1} p_i^2 \left[(2^{n-1} - 1) p_i + 1 \right]^2. \end{aligned} \quad (2.5)$$

In later chapters, the ratio $\Pr(\Delta_i \mid R_{1:n}) / \Pr(\Delta_i \mid R_{n:1,1,\dots})$ is used so it is convenient to find the probability of obtaining n individuals with the same genotypes as found in

pattern Δ_i given that they are unrelated ($R_{n:1,1,\dots}$), and list the simplified expressions (see Section 2.3.4 for details). For type Δ_1

$$\Pr(\Delta_1 \text{ genotypes} \mid R_{n:1,1,\dots}) = \prod_{i=1}^n \Pr(A_i A_i) = p_i^{2n}. \quad (2.6)$$

The joint probabilities for the other family genotype patterns can be calculated in a similar way (see Appendix (A)) and are summarised below. For each case, $i, j, k, l = 1, 2, \dots, s, i \neq j \neq k \neq l, 0 < p_x < 1, r_i \neq 0$ and $\sum_i r_i = n$.

Family Pattern Δ_2 The sample consists of n identical heterozygotes, denoted $nA_i A_j$.

$$\Pr(\Delta_2 \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i p_j (2^{n-1} (2^{n-1} - 1) p_i p_j + (2^{n-1} - 1) p_i + (2^{n-1} - 1) p_j + 1) \quad (2.7)$$

$$\Pr(\Delta_2 \mid R_{n:1,1,\dots}) = (2p_i p_j)^n \quad (2.8)$$

Family Pattern Δ_3 The sample has $r_1 A_i A_i$ homozygotes and $r_2 A_j A_j$ homozygotes.

$$\Pr(\Delta_3 \mid R_{1:n}) = \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j^2 \quad (2.9)$$

$$\Pr(\Delta_3 \mid R_{n:1,1,\dots}) = p_i^{2r_1} p_j^{2r_2} \quad (2.10)$$

Family Pattern Δ_4 Here there are r_1 homozygotes and r_2 heterozygotes, such that the heterozygous individuals have one allele in common with the homozygotes; $r_1 A_i A_i$ and $r_2 A_i A_j$.

$$\Pr(\Delta_4 \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j ((2^{n-1} - 1) p_i + (2^{r_2-1} - 1) p_j + 1) \quad (2.11)$$

$$\Pr(\Delta_4 \mid R_{n:1,1,\dots}) = p_i^{2r_1} (2p_i p_j)^{r_2} \quad (2.12)$$

Family Pattern Δ_5 This type of sample has two kinds of homozygote, with heterozygotes that have the same two alleles as the homozygotes; $r_1 A_i A_i, r_2 A_i A_j$ and $r_3 A_j A_j$.

$$\Pr(\Delta_5 \mid R_{1:n}) = 2^{r_2} \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j^2 \quad (2.13)$$

$$\Pr(\Delta_5 \mid R_{n:1,1,\dots}) = p_i^{2r_1} (2p_i p_j)^{r_2} p_j^{2r_3} \quad (2.14)$$

Family Pattern Δ_6 The sample has r_1 homozygotes and r_2 heterozygotes with no alleles in common; $r_1 A_i A_i$ and $r_2 A_j A_k$.

$$\Pr(\Delta_6 \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j p_k \quad (2.15)$$

$$\Pr(\Delta_6 \mid R_{n:1,1,\dots}) = p_i^{2r_1} (2p_j p_k)^{r_2} \quad (2.16)$$

Family Pattern Δ_7 There are two sorts of heterozygote in this type of sample, they have one allele in common; $r_1 A_i A_j$ and $r_2 A_i A_k$.

$$\Pr(\Delta_7 \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i p_j p_k (2^{n-1} p_i + 1) \quad (2.17)$$

$$\Pr(\Delta_7 \mid R_{n:1,1,\dots}) = 2^n (p_i p_j)^{r_1} (p_i p_k)^{r_2} \quad (2.18)$$

Family Pattern Δ_8 The sample has one sort of homozygote and two types of heterozygote, with one allele common to all; $r_1 A_i A_i$, $r_2 A_i A_j$ and $r_3 A_i A_k$.

$$\Pr(\Delta_8 \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j p_k \quad (2.19)$$

$$\Pr(\Delta_8 \mid R_{n:1,1,\dots}) = p_i^{2r_1} (2p_i p_j)^{r_2} (2p_i p_k)^{r_3} \quad (2.20)$$

Family Pattern Δ_9 There are two types of heterozygote, with no alleles in common and a homozygote with an allele in common with one heterozygote; $r_1 A_i A_i$, $r_2 A_i A_j$ and $r_3 A_j A_k$.

$$\Pr(\Delta_9 \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j p_k \quad (2.21)$$

$$\Pr(\Delta_9 \mid R_{n:1,1,\dots}) = p_i^{2r_1} (2p_i p_j)^{r_2} (2p_j p_k)^{r_3} \quad (2.22)$$

Family Pattern Δ_{10} A sample with this pattern has one sort of homozygote, two sorts of heterozygote with an allele in common with the homozygote, and a third type of heterozygote that has an allele in common with both other heterozygotes (but not the homozygote); $r_1 A_i A_i$, $r_2 A_i A_j$, $r_3 A_i A_k$ and $r_4 A_j A_k$.

$$\Pr(\Delta_{10} \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j p_k \quad (2.23)$$

$$\Pr(\Delta_{10} \mid R_{n:1,1,\dots}) = p_i^{2r_1} (2p_i p_j)^{r_2} (2p_i p_k)^{r_3} (2p_j p_k)^{r_4} \quad (2.24)$$

Family Pattern Δ_{11} This type of sample is like family pattern Δ_{10} , but does not have any homozygotes; $r_1A_iA_j$, $r_2A_iA_k$ and $r_3A_jA_k$.

$$\Pr(\Delta_{11} \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i p_j p_k (p_i + p_j + p_k) \quad (2.25)$$

$$\Pr(\Delta_{11} \mid R_{n:1,1,\dots}) = 2^n (p_i p_j)^{r_1} (p_i p_k)^{r_2} (p_j p_k)^{r_3} \quad (2.26)$$

Family Pattern Δ_{12} There are four different alleles present in this sample; there are three different heterozygotes, each has an allele in common with one other heterozygote; $r_1A_iA_k$, $r_2A_iA_l$ and $r_3A_jA_k$.

$$\Pr(\Delta_{12} \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i p_j p_k p_l \quad (2.27)$$

$$\Pr(\Delta_{12} \mid R_{n:1,1,\dots}) = 2^n (p_i p_k)^{r_1} (p_i p_l)^{r_2} (p_j p_k)^{r_3} \quad (2.28)$$

Family Pattern Δ_{13} This type of sample has two different heterozygotes that have no alleles in common; $r_1A_iA_k$ and $r_2A_jA_l$.

$$\Pr(\Delta_{13} \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i p_j p_k p_l \quad (2.29)$$

$$\Pr(\Delta_{13} \mid R_{n:1,1,\dots}) = 2^n (p_i p_k)^{r_1} (p_j p_l)^{r_2} \quad (2.30)$$

Family Pattern Δ_{14} There are four different types of heterozygote in this family pattern; $r_1A_iA_k$, $r_2A_iA_l$, $r_3A_jA_k$ and $r_4A_jA_l$.

$$\Pr(\Delta_{14} \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i p_j p_k p_l \quad (2.31)$$

$$\Pr(\Delta_{14} \mid R_{n:1,1,\dots}) = 2^n (p_i p_k)^{r_1} (p_i p_l)^{r_2} (p_j p_k)^{r_3} (p_j p_l)^{r_4} \quad (2.32)$$

These expressions were derived independently, but are also given in a paper by Elandt-Johnson [1971a].

2.3.4 A Sample of F families

The previous section shows how $\Pr(\Delta_i \mid R_{1:n})$ can readily be calculated, and provide simple expressions for the 14 general patterns that will occur in samples. The extension to relationships that include more than one family is straightforward.

For any N , there are a known set of relationships (Section 2.2) under the simple model described. Taking each relationship structure in turn, the individuals in the sample are sorted into families and every grouping that matches one of the 14 patterns (Table 2.3) is called a *valid* family. As the genotypes of individuals with different parents are independent, the overall probability for the relationship is the product of the probabilities for each constituent family. If any family is found to be incompatible on the basis of the genotypes of the siblings (at any locus), i.e. the genotypes do not match any family pattern, then any relationship that includes this family is said to be *excluded* and is discounted from further analyses. The effect of exclusion on the inference of relationship is considered further in Chapter 4.

For example, consider a sample of six individuals and the relationship $R_{3:3,2,1}$ as

$$\{X_2, X_3, X_6\}\{X_1, X_4\}\{X_5\}.$$

Suppose these individuals have the following genotypes (γ_i) for locus 1, which has two codominant alleles A_1 and A_2 ,

$$\gamma_1 = A_1A_1, \gamma_2 = A_1A_1, \gamma_3 = A_1A_2, \gamma_4 = A_2A_2, \gamma_5 = A_1A_2, \gamma_6 = A_2A_2.$$

Taking each family separately:

- Family 1, $\{X_2, X_3, X_6\}$ or $\{A_1A_1, A_1A_2, A_2A_2\}$. These genotypes fit family pattern Δ_5 with $r_1 = r_2 = r_3 = 1$ and $n = 3$. From Expression (2.13)

$$\Pr(\Delta_5 \mid R_{1:3}) = 2^{r_2} \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j^2 = \frac{1}{8} p_1^2 p_2^2. \quad (2.33)$$

- Family 2, $\{X_1, X_4\}$ or $\{A_1A_1, A_2A_2\}$, which is family pattern Δ_3 with $r_1 = r_2 = 1$ and $n = 2$. From Expression (2.9),

$$\Pr(\Delta_3 \mid R_{1:2}) = \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j^2 = \frac{1}{4} p_1^2 p_2^2. \quad (2.34)$$

- Family 3, $\{X_5\}$ or $\{A_1A_2\}$ which is family pattern Δ_2 with $n=1$. From Expression (2.7),

$$\Pr(\Delta_2 \mid R_{1:1}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i p_j (2^{n-1} (2^{n-1} - 1) p_i p_j) = 2 p_1 p_2, \quad (2.35)$$

which is the population frequency of A_1A_2 .

Hence, the overall probability for the relationship $\{X_2, X_3, X_6\}\{X_1, X_4\}\{X_5\}$ is equal to the product of Expressions (2.33) – (2.35), namely

$$\left(\frac{1}{8}p_1^2p_2^2\right) \left(\frac{1}{4}p_1^2p_2^2\right) (2p_1p_2) = \frac{1}{16}p_1^5p_2^5. \quad (2.36)$$

In this way it is possible to calculate the likelihood for any relationship by combining the appropriate expressions from Section 2.3.3.

It should also be noted that if $n = 1$ in Expression (2.3), then this expression of course becomes equal to $\Pr(\gamma)$, the population frequency of that genotype. The relationship $R_{n:1,1,\dots}$ is the same as considering n families each with one member (where each family must fit either pattern Δ_1 or Δ_2), so that

$$\Pr(\gamma_1\gamma_2\cdots\gamma_n \mid R_{n:1,1,\dots}) = \prod_{i=1}^n \left[\sum_{u=1}^g \sum_{\substack{v=1 \\ u \leq v}}^g \alpha_{u,v} \Pr(\gamma_i \mid \gamma_u \times \gamma_v) \right] = \prod_{i=1}^n \Pr(\gamma_i). \quad (2.37)$$

2.3.5 Multiple Loci

The genotypes at different loci are independent if the loci are unlinked. So if there are m loci, the likelihood for relationship R_* is simply given by

$$\Pr(\Gamma_1, \Gamma_2, \dots, \Gamma_n \mid R_*) = \prod_{j=1}^m \Pr(\gamma_{1j}, \gamma_{2j}, \dots, \gamma_{nj} \mid R_*). \quad (2.38)$$

If the loci are independent, then taking each locus separately gives an estimate of the relationship between individuals in the sample. These could be used to investigate the variability of the estimates.

2.4 Likelihood Ratios

Inferences regarding the degree of relatedness within the group of individuals are based on these likelihoods. The actual value of the likelihood is not important, however; it is the *relative* magnitude of the likelihoods that is of interest. Comparisons are therefore made by calculating a likelihood ratio for given hypotheses. Rather than calculate all possible pairwise ratios, each hypothesis is compared with one particular hypothesis. This hypothesis may be chosen arbitrarily, but throughout this study the relationship in which all members of the sample are unrelated is used as the basic structure against

which all others are measured. This is the logical choice as this relationship occurs for a sample of any size and can never be excluded on the basis of the genotypes, i.e. the likelihood for $R_{N:1,1,\dots}$ is always greater than zero (although it may be very small). Hence, the likelihood ratio for a relationship, $R_{F:n_1,n_2,\dots,n_F}$, is given by

$$\frac{\Pr(\Gamma_1, \Gamma_2, \dots, \Gamma_N \mid R_{F:n_1,n_2,\dots,n_F})}{\Pr(\Gamma_1, \Gamma_2, \dots, \Gamma_N \mid R_{N:1,1,\dots})} \quad (2.39)$$

where $N = \sum_{i=1}^F n_i$.

The likelihood ratio for each relationship is calculated, and the set of ratios sorted into descending order of magnitude. The relative positions of the relationship hypotheses in this ranking are used to make inferences about the relatedness of the sample of individuals. The likelihood for a relationship may also be presented as the natural logarithm of the probabilities shown in Expression (2.3); these log-likelihoods are called the supports by Edwards [1972] and Thompson [1975]. The natural logarithm of the likelihood ratios, as shown in Expression (2.39), are similar to the support differences used by Thompson.

2.5 Ordering of Individuals and Families

The likelihood for a relationship is defined as the probability of obtaining an ordered set of individuals with the observed genotypes, given this particular relationship. For example, the likelihood for the relationship $R_{3:3,2,1}$ as $\{X_2, X_3, X_6\}\{X_1, X_4\}\{X_5\}$ is shown to be equal to $1/16(p_1^5 p_2^5)$ in Section 2.3.4. This is also the likelihood for the relationship $R_{3:3,2,1}$ as $\{X_5\}\{X_2, X_3, X_6\}\{X_1, X_4\}$, as $\{X_2, X_6, X_3\}\{X_5\}\{X_1, X_4\}$, or as $\{X_4, X_1\}\{X_3, X_2, X_6\}\{X_5\}$ since these describe the same genetic relationship but simply consider the individuals in a different order within families or a different order of families.

Consider now the likelihood for a relationship for an *unordered* set of individuals. If L represents the likelihood for a relationship for an ordered set of individuals, then clearly the likelihood for an unordered set is L/β where β is the number of ways of ordering the individuals when describing the same relationship.

Suppose the sample consists of three individuals labelled X_1, X_2 and X_3 . There are $3!=6$ ordered sets, namely $X_1X_2X_3$, $X_1X_3X_2$, $X_2X_1X_3$, $X_2X_3X_1$, $X_3X_1X_2$ and

$X_3X_2X_1$, and five distinct genetic relationship structures:

1. "All three full sibs" ($R_{1:3}$)
2. "Two full sibs and the third one unrelated" ($R_{2:2,1}$)
3. "Two full sibs and the second one unrelated" ($R_{2:2,1}$)
4. "Two full sibs and the first one unrelated" ($R_{2:2,1}$)
5. "All three unrelated" ($R_{3:1,1,1}$).

Hence, for a sample of three individuals, the following 30 equally likely events can be defined:

$\{X_1X_2X_3\}$	$\{X_1X_2\}\{X_3\}$	$\{X_1\{X_2\}X_3\}$	$\{X_1\}\{X_2X_3\}$	$\{X_1\}\{X_2\}\{X_3\}$
$\{X_1X_3X_2\}$	$\{X_1X_3\}\{X_2\}$	$\{X_1\{X_3\}X_2\}$	$\{X_1\}\{X_3X_2\}$	$\{X_1\}\{X_3\}\{X_2\}$
$\{X_2X_1X_3\}$	$\{X_2X_1\}\{X_3\}$	$\{X_2\{X_1\}X_3\}$	$\{X_2\}\{X_1X_3\}$	$\{X_2\}\{X_1\}\{X_3\}$
$\{X_2X_3X_1\}$	$\{X_2X_3\}\{X_1\}$	$\{X_2\{X_3\}X_1\}$	$\{X_2\}\{X_3X_1\}$	$\{X_2\}\{X_3\}\{X_1\}$
$\{X_3X_1X_2\}$	$\{X_3X_1\}\{X_2\}$	$\{X_3\{X_1\}X_2\}$	$\{X_3\}\{X_1X_2\}$	$\{X_3\}\{X_1\}\{X_2\}$
$\{X_3X_2X_1\}$	$\{X_3X_2\}\{X_1\}$	$\{X_3\{X_2\}X_1\}$	$\{X_3\}\{X_2X_1\}$	$\{X_3\}\{X_2\}\{X_1\}$

If the above sets are rewritten in the standard notation, it can be seen that there are $3! = 6$ permutations for each of the five structures 1: $\{X_1, X_2, X_3\}$, 2: $\{X_1, X_2\}\{X_3\}$, 3: $\{X_1, X_3\}\{X_2\}$, 4: $\{X_2, X_3\}\{X_1\}$ and 5: $\{X_1\}\{X_2\}\{X_3\}$. The coefficient β , the number of ways of ordering N individuals to describe the same basic relationship, is $N!$ for any N . Therefore, when calculating the likelihood ratios, the ordering coefficient cancels and can be omitted. The likelihood ratio for two relationships will be the same regardless of whether the individuals are considered as ordered or unordered.

The Simulation and Calculation Programs

Throughout Chapter 3, theoretical results are verified using computer simulation programs, whilst in Chapter 4 conclusions are based entirely on results obtained from the same programs. The main set of programs used to generate and analyse genotype data are described in Appendix B.1.

2.6 Summary

This chapter describes the relationship structures that are to be considered for a set of individuals, and the basic calculations necessary to find the likelihood for each relationship. A set of 14 family patterns are described, showing the possible genotypes that may be found in a valid family. The likelihood for each of these families is calculated, and the method for combining these expressions for more complex relationship structures consisting of more than one family is described.

A simple ratio of these likelihoods is calculated for each relationship, and then these ratios are ranked in descending order of magnitude. The relationship with the highest ratio is ranked in first position, and is the relationship for which the probability of obtaining those genotypes observed in the sample is greatest.

There are different ways to interpret these results. The simplest approach is to find the first ranked relationship and conclude that this is the best estimate of the relatedness in the sample. This procedure is followed in Chapter 3 for a number of examples, and the accuracy of the results is discussed.

Alternatively, it is possible to examine the relative order of relationship structures in the list and to identify possible trends. Inferences of relatedness based on these trends, taking account of the range of likelihood ratios, are discussed in Chapter 4.

Chapter 3

First Rank Inferences

3.1 Introduction

The number and type of relationship structures to be considered for a sample of N individuals are described in the previous chapter. The procedure for the calculation of the likelihood and likelihood ratio for each relationship hypothesis is also given.

The simplest way of interpreting these likelihood ratios is to rank the hypotheses such that the largest ratio is placed first, and then take the first ranked relationship as the best estimate of the structure within the sample.

In order to illustrate this method of analysis, the first part of this chapter considers in detail the simple case where the sample consists of just two individuals. In Section 3.2, details are given for the calculations involved and for the inferences that are made for each possible sample. This is illustrated using data for a single locus with s codominant alleles. Having found the inference of relationship that is made for each type of sample, it is then possible to calculate a “probability of correct inference”; this is shown in Section 3.2.2. Complete theoretical results are found for this simple example and the results are confirmed empirically using computer simulation experiments. In Section 3.2.2, the analysis is extended first to two loci and then to many multi-allele independent loci, using computer simulation. The change in the probability of correct inference with increasing information is examined.

The second part of this chapter shows the results expected and those obtained from simulation experiments for samples of more than two individuals. For each example,

the probability of correct inference of relationship is found for the case where the inference is based on the hypothesis ranked first. Samples with three individuals are examined briefly to illustrate those properties of the analysis not exhibited by two individuals – namely exclusion and a probability of correct inference equal to zero for some examples. Samples of four, five and six individuals are used to show typical results with the increasingly complex pattern of inference for larger sample sizes.

Finally, in Section 3.5 the general pattern of inference based on the first rank is summarised.

3.2 Details for a Sample of Two Individuals

Consider the simple case where two individuals are chosen for analysis, for example, two fledglings from a nest. For this analysis, it is assumed that these two individuals could be related in one of two ways; they are either full sibs, $R_{1:2}$, or unrelated, $R_{2:1,1}$. In order to infer the relationship, the probability of obtaining the observed genotypes is calculated and the relationship that gives the higher likelihood is selected.

As there are only two relationships, in the following sections the standard notation for each relationship ($R_{1:2}$ and $R_{2:1,1}$) is abbreviated to R_1 and R_2 , the subscript referring to the number of hypothetical families.

3.2.1 The Likelihood for Each Relationship

First consider the case where the two individuals in the sample are assumed to be full sibs. The likelihood for this relationship is given by

$$\Pr(\Gamma_1, \Gamma_2 \mid R_1) = \prod_{j=1}^m \Pr(\gamma_{1j}, \gamma_{2j} \mid R_1) \quad (3.1)$$

where m is the number of independent loci (from Expression (2.38)). For a locus with s alleles there are $\frac{1}{2}s(s+1) = g$ different genotypes, and therefore g^2 different ordered pairs of genotypes. Each of these pairs can be represented by one of seven general pairs (without regard to order) depending on the number of alleles the two genotypes have in common (Table 3.1). Furthermore, each of these pairs corresponds to one of the family

Family Pattern	$\Pr(\Delta_i R_1)$	$\Pr(\Delta_i R_2)$
$\Delta_1 - A_i A_i, A_i A_i$	$\frac{1}{4}p_i^2(1 + p_i)^2$	p_i^4
$\Delta_2 - A_i A_j, A_i A_j$	$\frac{1}{2}p_i p_j(2p_i p_j + p_i + p_j + 1)$	$4p_i^2 p_j^2$
$\Delta_3 - A_i A_i, A_j A_j$	$\frac{1}{4}p_i^2 p_j^2$	$p_i^2 p_j^2$
$\Delta_4 - A_i A_i, A_i A_j$	$\frac{1}{2}p_i^2 p_j(1 + p_i)$	$2p_i^3 p_j$
$\Delta_6 - A_i A_i, A_j A_k$	$\frac{1}{2}p_i^2 p_j p_k$	$2p_i^2 p_j p_k$
$\Delta_7 - A_i A_j, A_i A_k$	$\frac{1}{2}p_i p_j p_k(2p_i + 1)$	$4p_i^2 p_j p_k$
$\Delta_{13} - A_i A_k, A_j A_l$	$\frac{1}{2}p_i p_j p_k p_l$	$4p_i p_j p_k p_l$

Table 3.1: The seven possible (unordered) pairs of genotypes for a single locus with s alleles, $i \neq j \neq k \neq l$, and the corresponding probability of obtaining a sample with these genotypes given R_1 or R_2 .

patterns described in Chapter 2; for example, the pair of genotypes $A_i A_i, A_j A_j$ ($i \neq j$) corresponds to type Δ_3 . When there are only two individuals, for a single locus with s alleles, clearly only family patterns Δ_1 – Δ_4 , Δ_6 , Δ_7 and Δ_{13} will occur as the remaining patterns require three or more individuals. Table 3.1 shows these patterns and the simplified probability expressions found by substituting $N = 2$ into Expressions (2.5) to (2.30). So, if $\gamma_{11} = A_3 A_3$ and $\gamma_{21} = A_1 A_3$ for example, then the individuals are said to fit pattern Δ_4 and $\Pr(\Gamma_1, \Gamma_2 | R_1) = \frac{1}{2}p_3^2 p_1(1 + p_3)$. Note that when there are only two individuals every possible pair of genotypes will correspond to one of the family patterns.

Now consider the case where the two individuals in the sample are assumed to be unrelated. The likelihood for the relationship R_2 is given by

$$\Pr(\Gamma_1, \Gamma_2 | R_2) = \prod_{j=1}^m \Pr(\gamma_{1j}) \Pr(\gamma_{2j}) \quad (3.2)$$

(see Expression (2.37)). The probability of obtaining an ordered pair given that they are unrelated can be found for each general pair; these seven likelihoods are also shown in Table 3.1. For example, if $\gamma_{11} = A_1 A_5$ and $\gamma_{21} = A_2 A_3$ then the individuals are said to fit pattern Δ_{13} with $\Pr(\Gamma_1, \Gamma_2 | R_2) = 4p_1 p_2 p_3 p_5$.

Having found the likelihoods for the two relationships, the values can be compared. For $N = 2$, there is only one likelihood ratio to find and so the inference of relationship simply depends on whether this ratio is greater or less than one. At this stage, the

magnitude of the ratio is not considered, though this will be discussed in Chapter 4. It is straightforward to calculate the likelihood ratio and find the relationship that will be inferred for every possible sample that could be drawn from a population.

Family Pattern	Likelihood Ratio $\Pr(\Delta \mid R_1)/\Pr(\Delta \mid R_2)$	Inference
$\Delta_1 - A_i A_i, A_i A_i$	$(1 + p_i)^2 / (4p_i^2)$	R_1
$\Delta_2 - A_i A_j, A_i A_j$	$(2p_i p_j + p_i + p_j + 1) / (8p_i p_j)$	R_1
$\Delta_3 - A_i A_i, A_j A_j$	$1/4$	R_2
$\Delta_4 - A_i A_i, A_i A_j$	$(1 + p_i) / (4p_i)$	R_1 if $p_i < 1/3$, R_2 if $p_i > 1/3$, undefined if $p_i = 1/3$
$\Delta_6 - A_i A_i, A_j A_k$	$1/4$	R_2
$\Delta_7 - A_i A_j, A_i A_k$	$(2p_i + 1) / (8p_i)$	R_1 if $p_i < 1/6$, R_2 if $p_i > 1/6$, undefined if $p_i = 1/6$
$\Delta_{13} - A_i A_k, A_j A_l$	$1/8$	R_2

Note that these ratios are either constant or depend only on the frequency of an allele held in common by both individuals. If the two individuals have no alleles in common at the locus (Δ_3 , Δ_6 and Δ_{13}) then the likelihood ratio is constant and less than one and so they will always be inferred as unrelated regardless of the frequency of any allele. Alternatively, if the individuals have identical genotypes (Δ_1 and Δ_2) then the likelihood ratio is greater than one for any set of allele frequencies, p , and so they will always be inferred as sibs. For the remaining two patterns, Δ_4 and Δ_7 , the likelihood ratio and therefore the inference of relationship, depends on the frequency of the allele that the pair have in common. For Δ_4 , i.e. $(A_i A_i, A_i A_j)$, the ratio is greater than one if $p_i < 1/3$, and so the individuals are inferred as sibs only if the shared allele has low frequency. For Δ_7 , the shared allele must have frequency less than $1/6$ for the individuals to be inferred as sibs. If the frequency of the shared allele is exactly $1/3$ or $1/6$, then the likelihood ratios for samples of type Δ_4 or Δ_7 , respectively, are equal to one; thus no inference can be made.

These results are consistent with those expected intuitively; closely related individuals have identical genotypes more often than unrelated ones and rare alleles are more likely to be identical by descent. The inference of relationship made for any given pair

of individuals is clearly unaffected by the number of alleles at the locus and depends on allele frequencies for some samples only.

Suppose there are only two alleles or the data are combined such that 'A₁' represents the most frequent allele and 'A₂' represents all other alleles (hence $p_1 + p_2 = 1$). Family patterns Δ_6 , Δ_7 and Δ_{13} can now be discounted and the likelihoods can be further simplified as shown below for each relationship.

Pattern	$\Pr(\Delta \mid R_1)$	$\Pr(\Delta \mid R_2)$
Δ_1	$\frac{1}{4}p_i^2(1 + p_i)^2$	p_i^4
Δ_2	$p_i - 2p_i^3 + p_i^4$	$4p_i^2(1 - p_i)^2$
Δ_3	$\frac{1}{4}p_i^2(1 - p_i)^2$	$p_i^2(1 - p_i)^2$
Δ_4	$\frac{1}{2}p_i^2(1 - p_i^2)$	$2p_i^3(1 - p_i)$
else $\Pr(\Delta) = 0$ ($0 < p_i < 1$).		

Figure 3.1 shows these likelihoods for values of p_i from 0 to 1, and Figure 3.2 shows the log ratios as $\ln(\Pr(\Delta_i \mid R_1)) - \ln(\Pr(\Delta_i \mid R_2))$ for Δ_{1-4} . Note that although there are three genotypes and nine possible ordered pairs of genotypes for a locus with two codominant alleles, these belong to only four general patterns for p_i and p_j .

Throughout, it is assumed that the frequency of each allele in the general population is known; thus, the likelihood for each relationship structure can be evaluated. If the population allele frequencies are not known, then finding the inference of relationship for each sample type is more complex. For example, consider a sample of two individuals whose genotypes are A_iA_i and A_jA_j . These individuals represent a sample of type Δ_3 ; Figure 3.1 shows that the likelihood for each relationship is maximised when $p_i = 1/2$, and that the likelihood for relationship R_2 has maximum value. Thus a Δ_3 sample is inferred as R_2 . Similarly, if the allele frequencies are not known, Δ_2 samples are inferred as R_1 and Δ_4 samples are inferred as R_2 . For samples of type Δ_1 , no inference can be made as the likelihood is maximised when $p_i = 1$ for both relationship structures, and both then have the same likelihood. The maximum likelihood for each relationship could be found in this way, and inferences of relationship made for any sample (see Chapter 7).

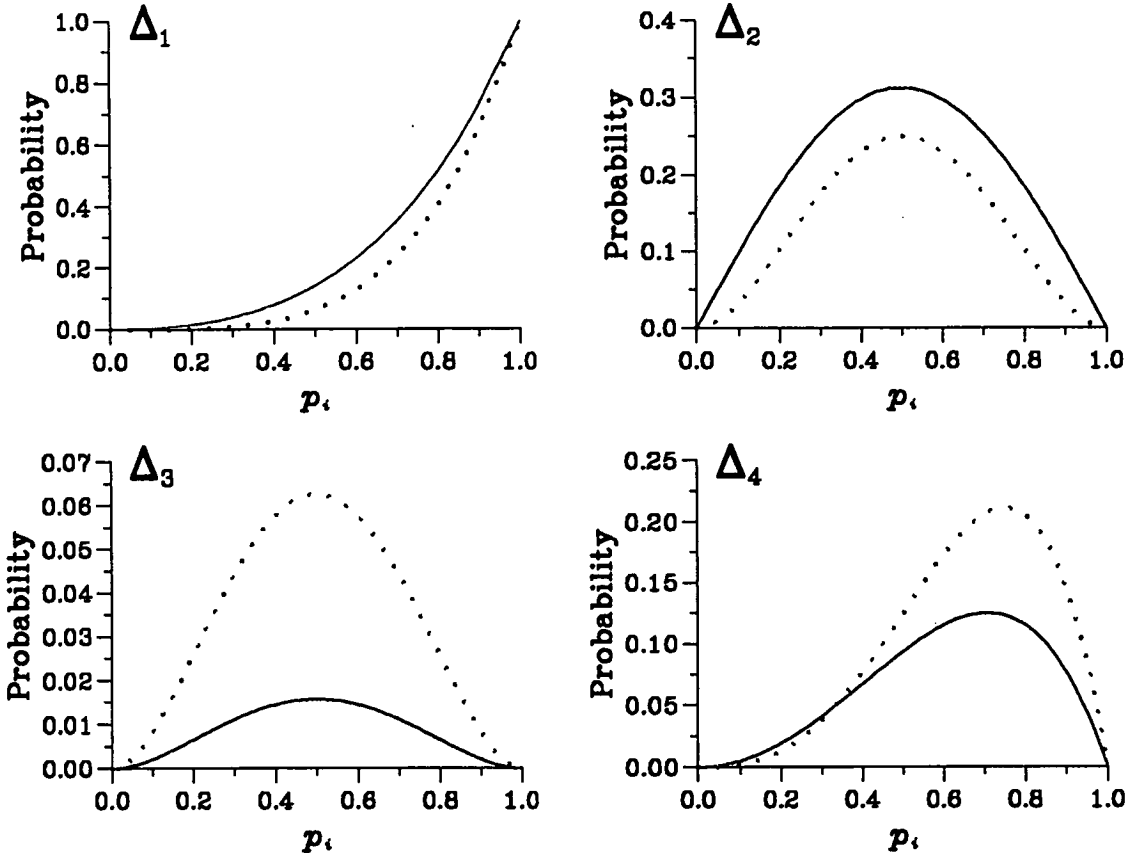


Figure 3.1: The probability of obtaining a sample of two individuals with genotypes fitting pattern Δ_i . The solid and dotted lines represent the probability of the data given R_1 and given R_2 , respectively, using data from one locus that has two alleles. Note that each graph is drawn to a different scale on the vertical axis.

3.2.2 The Probability of Correct Inference

The frequency with which a particular set of genotypes occurs depends on allele frequency and on the true relationship between the sampled individuals. For example, a pair of identical homozygotes A_1A_1, A_1A_1 is expected to occur with frequency p_1^4 if the individuals are unrelated but expected to occur with frequency $\frac{1}{4}p_1^2(1+p_1)^2$ if sampled as a pair from a population of full sibling pairs. In either case the same relationship is inferred; thus, if the pair *are* full siblings the correct inference is made whereas if they are not closely related, the wrong conclusion is drawn. From this it is possible to calculate a “probability of correct inference”, that is, the probability of inferring

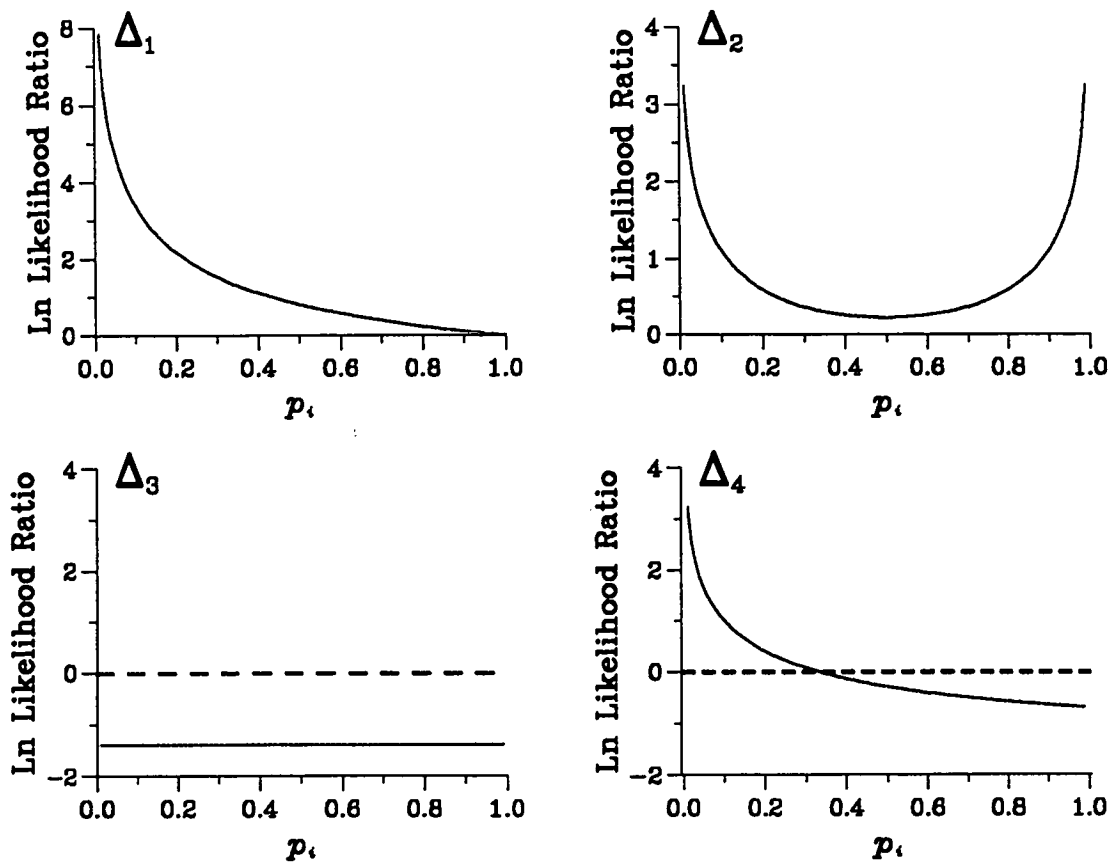


Figure 3.2: The log-likelihood ratio, $\ln(\Pr(\Delta_i | R_1)) - \ln(\Pr(\Delta_i | R_2))$ for each of the four distinct patterns seen with a sample of size two and data from one two-allele locus. The dashed line shows a likelihood ratio equal to one and hence the value around which the inference changes.

relationship R_* given that R_* is the true relationship:

$$\begin{aligned} \Pr(\text{Correct Inference}) &= \Pr(\text{sampling individuals related as } R_* \text{ and inferring } R_*) \\ &= \sum_{\text{Relationships}} \Pr(\text{inferring } R_* | R_*) \Pr(R_*) \end{aligned} \quad (3.3)$$

Suppose there are a given number of individuals and a set of relationship hypotheses for this number. Each hypothesis has a prior likelihood based on known mating patterns and family size distributions within the population. In field experiments the prior probability of a relationship, e.g. the probability of sampling two full siblings ($\Pr(R_{1:2})$), is generally unknown. However, the probability of correct inference can be readily investigated by using simulation experiments. By using the simulation program (Appendix B.1) to generate the genotypes of individuals with known relatedness, $\Pr(R_*)$ in Expression (3.3) becomes equal to one. Hence, by analysing many replicate

samples, the proportion of samples correctly inferred is an estimate of the probability of correct inference.

First, this probability is investigated for a sample of two individuals for a single locus with s alleles and then this is extended to multiple loci. The theoretical results are found in each case and compared with those obtained by simulation in order to check the programs that are used for more complex examples in subsequent chapters.

A Single Locus

The number of different pairs of genotypes fitting each family pattern depends on the number of alleles. The frequency with which each type occurs depends on the allele frequencies and on the true relationship between the individuals. For example, if $s = 3$ then there will be $s(s-1)(s-2) = 6$ pairs of genotypes that fit pattern Δ_6 ; A_1A_1 and A_2A_3 , A_2A_3 and A_1A_1 , A_2A_2 and A_1A_3 , A_1A_3 and A_2A_2 , A_3A_3 and A_1A_2 , and A_1A_2 and A_3A_3 . A pair fitting pattern Δ_6 will occur with frequency $\frac{1}{2}p_i^2p_jp_k$ if sampled from full sibling pairs, and with frequency $2p_i^2p_jp_k$ if sampled from unrelated individuals. All these samples will be inferred as unrelated regardless of the allele frequency distribution. Noting the inference of relationship that will be made for each type of sample (shown above), for any allele frequency distribution and any number of alleles the probability of inferring R_1 is equal to

$$\Pr(\Delta_1) + \Pr(\Delta_2) + \left\{ \Pr(\Delta_4) \text{ if } p_i < \frac{1}{3} \right\} + \left\{ \Pr(\Delta_7) \text{ if } p_i < \frac{1}{6} \right\}, \quad (3.4)$$

and the probability of inferring R_2 is

$$\Pr(\Delta_3) + \Pr(\Delta_6) + \Pr(\Delta_{13}) + \left\{ \Pr(\Delta_4) \text{ if } p_i > \frac{1}{3} \right\} + \left\{ \Pr(\Delta_7) \text{ if } p_i > \frac{1}{6} \right\}. \quad (3.5)$$

Here, $\Pr(\Delta_1)$, for example, is the probability of drawing a pair of individuals whose genotypes fit the family pattern Δ_1 . This of course depends on the relatedness of the population. Note that if $p_i = 1/3$, then neither R_1 nor R_2 is inferred for a Δ_4 sample as the likelihood ratio is equal to one. To find the probability of *correctly* inferring relationship R_* , the appropriate expressions for $\Pr(\Delta_i)$ are substituted in Expressions (3.4) and (3.5).

For a single locus with two alleles, $\Pr(\Delta_7)$ and $\Pr(\Delta_{13})$ equal zero and the probability of correct inference is straightforward; this is shown in Figure 3.3 for a range of values of

p_1 from 0 to 1. At extreme gene frequencies, most samples drawn from the population will consist of pairs of individuals homozygous for the common allele and will therefore be inferred as siblings. The graph is symmetrical about $p_i = 0.5$ as there are only two alleles and $\sum p_i = 1$ and the discontinuity in gradient occurs at $p_i = 1/3$ where there is a change of inference for some samples. Observed results from simulation agree well with those expected.

When there are more than two alleles, the number of samples fitting each pattern increases; the calculation of $\Pr(\Delta_i)$ depends on the frequency of each allele and so is more complex. For simplicity, let each of the s alleles have equal frequency, i.e. $p_i = 1/s$. The likelihoods for the relationships R_1 and R_2 can be further simplified thus,

Family Pattern	Number	$\Pr(\Delta R_1)$	$\Pr(\Delta R_2)$
$\Delta_1 \quad A_i A_i, A_i A_i$	s	$\frac{1}{4}p^2(1+p)^2$	p^4
$\Delta_2 \quad A_i A_j, A_i A_j$	$\frac{1}{2}s(s-1)$	$\frac{1}{2}p^2(2p^2 + 2p + 1)$	$4p^4$
$\Delta_3 \quad A_i A_i, A_j A_j$	$s(s-1)$	$\frac{1}{4}p^4$	p^4
$\Delta_4 \quad A_i A_i, A_i A_j$	$2s(s-1)$	$\frac{1}{2}p^3(1+p)$	$2p^4$
$\Delta_6 \quad A_i A_i, A_j A_k$	$s(s-1)(s-2)$	$\frac{1}{2}p^4$	$2p^4$
$\Delta_7 \quad A_i A_j, A_i A_k$	$s(s-1)(s-2)$	$\frac{1}{2}p^3(2p+1)$	$4p^4$
$\Delta_{13} \quad A_i A_k, A_j A_l$	$\frac{1}{4}s(s-1)(s-2)(s-3)$	$\frac{1}{2}p^4$	$4p^4$

So when drawing a pair of siblings from a population of full sibling pairs ($\Pr(R_1) = 1$), from Expression (3.4) there is a proportion

$$s \left(\frac{1}{4}p^2(1+p)^2 \right) + \left(\frac{1}{2}s(s-1) \right) \left(\frac{1}{2}p^2(2p^2 + 2p + 1) \right) \quad (3.6)$$

of samples always correctly inferred as siblings, a proportion

$$2s(s-1) \left(\frac{1}{2}p^3(1+p) \right) \quad (3.7)$$

correctly inferred as siblings if $p < 1/3$, that is when there are more than three alleles (in this example), and a proportion

$$s(s-1)(s-2) \left(\frac{1}{2}p^3(2p+1) \right) \quad (3.8)$$

correctly inferred as siblings if $p < 1/6$, i.e. $s > 6$.

Similarly, when drawing unrelated pairs from a large population ($\Pr(R_2) = 1$), expressions can be derived for the probability of correct inference given R_2 . In this

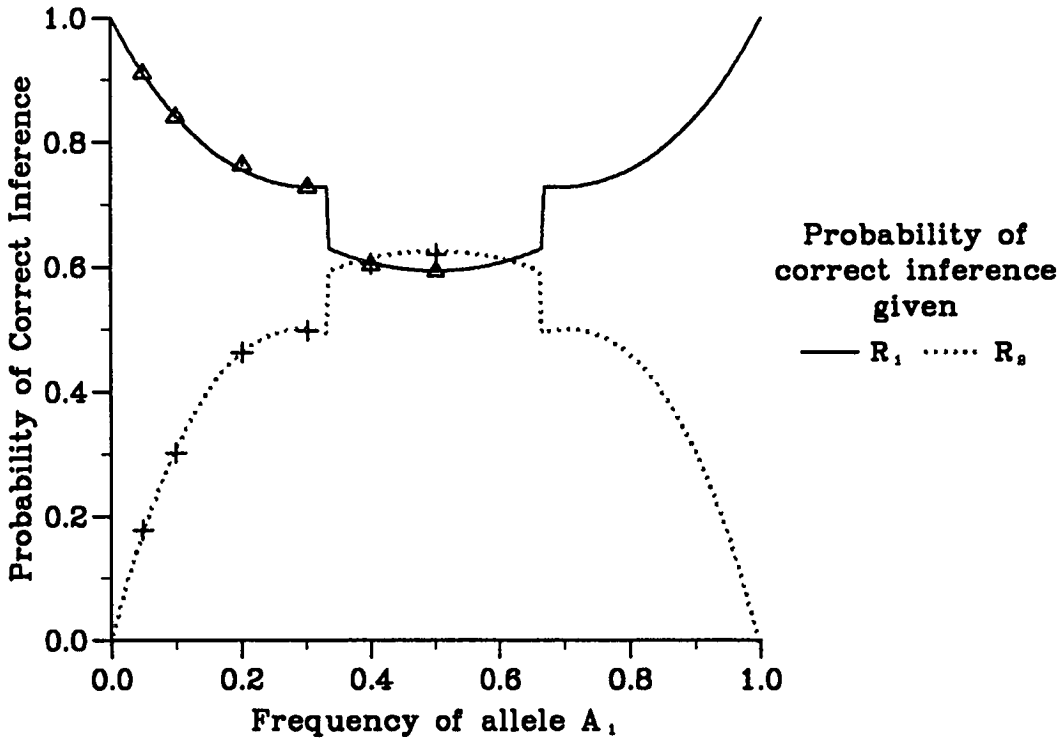


Figure 3.3: The probability of correct inference for two individuals and one locus with two alleles. Points show results from simulation experiments, each point represents the proportion of samples correctly inferred from 10,000 replicate trials.

case, from Expression (3.5) a proportion

$$s(s-1)p^4 + s(s-1)(s-2)(2p^4) + \frac{1}{4}s(s-1)(s-2)(s-3)4p^4 \quad (3.9)$$

of samples will always be correctly inferred as R_2 , plus

$$2s(s-1)2p^4 \quad (3.10)$$

correctly inferred as unrelated if $p > \frac{1}{3}$, i.e. $s < 3$, plus

$$s(s-1)(s-2)4p^4 \quad (3.11)$$

correctly inferred if $p > 1/6$, i.e. $2 < s < 6$. Note that when $s = 3$ or $s = 6$, and therefore $p = 1/3$ or $p = 1/6$, the ratio is exactly equal to one and no inference can be made for some samples.

The expected probability of correct inference and the observed results using simulation are shown in Figures 3.4a and 3.4b. There is a change of inference at $p = 1/3$ or

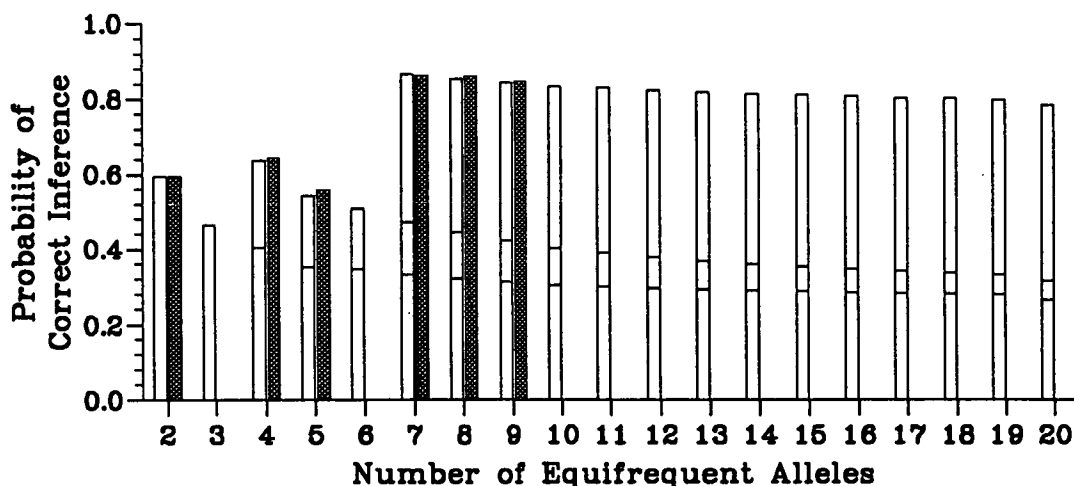


Figure 3.4a: The probability of correct inference given R_1 for $N = 2$, and one locus with s alleles where $p = 1/s$.

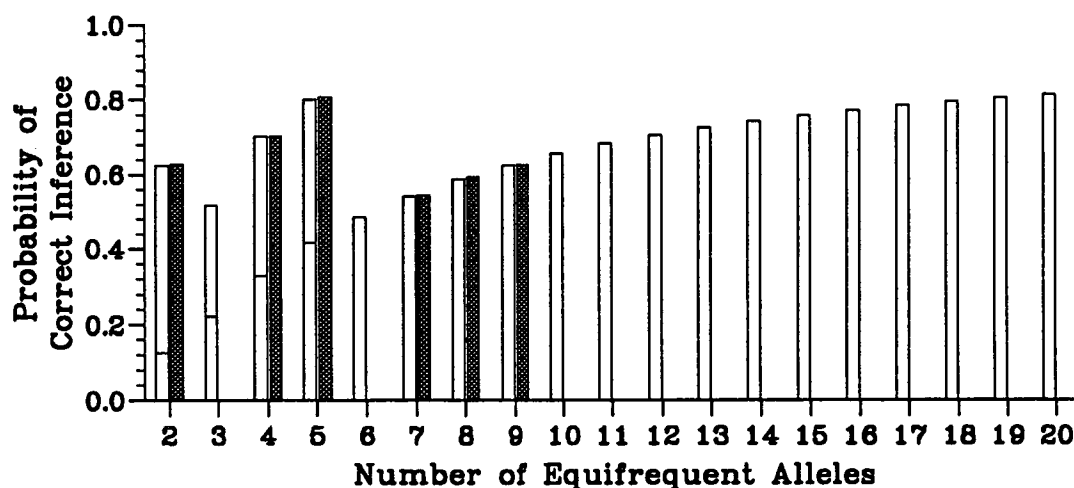


Figure 3.4b: The probability of correct inference given R_2 for $N = 2$, and one locus with s alleles where $p = 1/s$.

Open columns represent theoretical values and the divisions indicate the contribution from each group of family patterns (see text). Solid columns show results observed from simulation experiments.

$p = 1/6$ for some sample types, causing sudden changes in the pattern of inference. For this example, this depends on whether there are more or less than three or six alleles. When the two individuals are full siblings (Figure 3.4a), the frequency of type Δ_1 and Δ_2 samples decreases as the number of alleles increases (this is shown by the height of the lowest division in the open columns) and the probability of correct inference decreases accordingly. But, when there are more than three (equifrequent) alleles type Δ_4 samples are also correctly inferred (shown by the middle division in the open columns) and when s is greater than six, Δ_7 samples contribute to the total (the topmost division). Note that the maximum probability of correct inference is 0.86 when there are seven alleles each with frequency $1/7$; the probability of correct inference decreases to 0.78 for 20 alleles. When there are many alleles, fewer samples will have individuals with alleles in common and so fewer are inferred as siblings. For Figure 3.4b, the sampled pair of individuals are unrelated. When there are two equifrequent alleles, type Δ_3 (lower division in open columns) and Δ_4 (upper division) samples are correctly inferred. When there are three alleles, Δ_4 are no longer correctly inferred but Δ_7 can now be included. The probability of correct inference reaches 0.80 when $s = 5$, but drops when Δ_7 samples are no longer correctly inferred. As the alleles increase in number but decrease in frequency more samples will be of types Δ_3 , Δ_6 and Δ_{13} and so the probability of correct inference of R_2 rises, reaching 0.81 for 20 alleles.

The solid columns show the observed results from simulation experiments; the programs are described in Appendix B.1. The probability of inferring R_* given R_* is the true relationship is estimated from the proportion of samples correctly inferred as R_* . Generally, the observed results agree well with those expected. Observed results, however, are not accurate for $s = 3$ or $s = 6$ because the allele frequencies are, of course, entered as real numbers and therefore cannot be exactly equal to $1/s$ for the programs used. This only affects the results significantly when $s = 3$ or $s = 6$ as there is a switch of inference at $p = 1/3$ and $p = 1/6$ and the inference of relationship is sensitive to p_i . In practice it is unlikely that allele frequencies would be exactly $1/3$ or $1/6$ so this is not a significant problem when analysing real data.

Simulation results are given below for a number of examples in which the s alleles do not have equal frequency. For the following Tables, the alleles frequency distribution is given by $\underline{p}$ and, for example, the notation $(5(0.1), 0.5)$ is used to denote a locus with six

alleles, five alleles with frequency 0.1 and one allele with frequency 0.5. Each simulation run consists of 10,000 trials; the 95% confidence limits are shown for each result. The confidence limits for a binomial proportion are given by $\pm 1.96 \times \sqrt{(\hat{q}(1 - \hat{q})/10000)}$, where $\hat{q}$ is the proportion of samples correctly inferred.

1. This example shows the probability of correct inference for a single locus and a range of allele frequencies for $s = 3, 4, 6$ and 8. The probabilities are generally higher for the case where the two individuals are not related, except when one allele is very common. Generally, as the number of alleles increases, more samples are correctly inferred as unrelated.

<i>s</i>	<i>p</i>	Pr(Correct Inference)	
		given <i>R</i> ₁	given <i>R</i> ₂
3	2(0.1), 0.8	0.7347±0.0087	0.5218±0.0098
	0.1, 0.3, 0.6	0.6754±0.0092	0.6281±0.0095
4	3(0.1), 0.7	0.6775±0.0092	0.6546±0.0093
	0.1, 0.2, 0.3, 0.4	0.6039±0.0096	0.7361±0.0086
6	5(0.1), 0.5	0.6426±0.0094	0.7627±0.0083
	3(0.1), 2(0.2), 0.3	0.6570±0.0093	0.7234±0.0088
8	7(0.05), 0.65	0.6309±0.0095	0.7523±0.0085
	2(0.05), 3(0.1), 3(0.2)	0.6469±0.0094	0.7682±0.0083

2. Typical loci often have one common allele plus a number of rare alleles. In the following set of simulation experiments, one allele was kept at a constant frequency ($p_1=0.25, 0.5$ or 0.75) and the number of other alleles increased from 1 to 8, each with equal frequency ($p_2 \dots p_s = (1 - p_1)/(s - 1)$). As before, each set of simulations shows the proportion of samples for which the correct relationship is inferred from 10,000 replicate trials, and the 95% confidence limits.

<i>s</i>	<i>p</i> ₁ = 0.25	<i>p</i> ₁ = 0.5	<i>p</i> ₁ = 0.75
	given <i>R</i> ₁	given <i>R</i> ₁	given <i>R</i> ₁
2	0.7345±0.0087	0.5932±0.0096	0.7370±0.0086
3	0.5387±0.0098	0.6108±0.0096	0.7096±0.0089
4	0.6414±0.0094	0.6272±0.0095	0.6936±0.0090
5	0.5663±0.0097	0.6484±0.0094	0.6857±0.0091
6	0.8021±0.0078	0.6409±0.0096	0.6826±0.0091
7	0.7786±0.0081	0.6411±0.0094	0.6737±0.0092
8	0.7682±0.0083	0.6270±0.0095	0.6784±0.0092
9	0.7595±0.0084	0.6198±0.0095	0.6792±0.0091

For a full sib pair, when $p_1 = 0.75$ the other alleles have low frequency and increasing their number has little effect on the probability of correct inference which decreases as s increases. If $p_1 = 0.25$ or 0.5 , then the frequency of the other alleles affects the pattern of inference such that the probability of correct inference does not change smoothly with increasing numbers of alleles. This depends on whether the allele frequencies are greater or less than $1/3$ or $1/6$.

For unrelated pairs, the change in the probability of correct inference is again influenced by whether the alleles have frequency less than or greater than $1/3$ and $1/6$.

s	$p_1 = 0.25$ given R_2	$p_1 = 0.5$ given R_2	$p_1 = 0.75$ given R_2
2	0.5069±0.0098	0.6200±0.0095	0.4915±0.0098
3	0.7648±0.0083	0.6959±0.0090	0.5746±0.0097
4	0.7022±0.0090	0.7180±0.0088	0.6147±0.0095
5	0.7933±0.0079	0.7249±0.0088	0.6267±0.0095
6	0.5820±0.0097	0.7646±0.0083	0.6444±0.0094
7	0.6352±0.0094	0.7821±0.0081	0.6390±0.0094
8	0.6808±0.0091	0.8036±0.0078	0.6497±0.0094
9	0.7015±0.0090	0.8240±0.0075	0.6539±0.0093

As expected, when the number of alleles increases such that most are rare ($p < 0.1$), there is little change in the probability of correct inference. As the number of alleles increases, more samples will be of pattern type Δ_3 , Δ_6 and Δ_{13} in which the individuals have no alleles in common and hence are inferred as R_2 . Thus the probability of correct inference will tend to rise for R_2 , but decrease for R_1 , with increasing number of alleles. If one or two alleles are very common, however, then pattern types Δ_1 and Δ_2 will predominate and the probability of correct inference for R_1 will be high. In neither case will the probability of correct inference be equal to one as there will always be some samples incorrectly inferred.

Multiple Loci

In the previous sections, data from only one locus are considered. In most situations, however, data would be available from several loci; hence the analysis is now extended to include multiple independent loci.

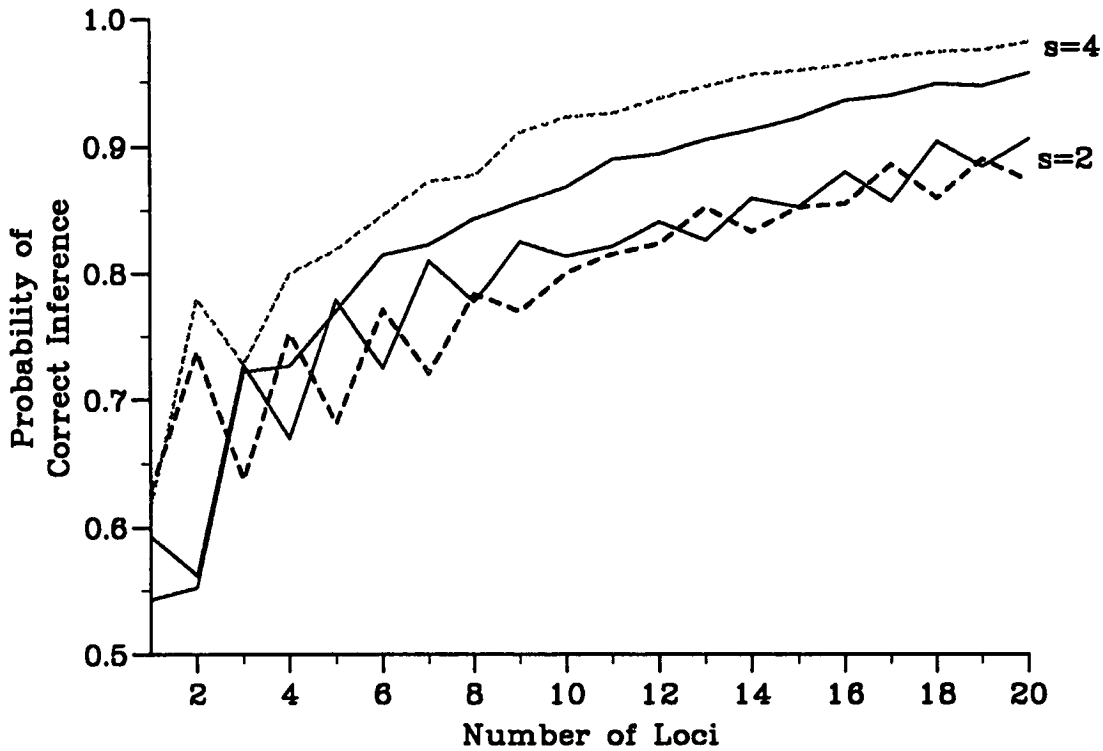


Figure 3.5: The estimated probability of correct inference from simulation experiments for a sample of two individuals with data from 1 to 20 loci, each locus has two or four alleles, each with equal frequency. The solid and dashed lines represent the probability of correct inference given R_1 and R_2 respectively.

Provided the loci are independent, it is straightforward to combine information. The overall probability is simply the product of the probabilities for each locus separately. Hence,

$$\Pr(\Gamma_1, \Gamma_2, \dots, \Gamma_N | R_*) = \prod_{j=1}^m \Pr(\gamma_{1j}, \gamma_{2j}, \dots, \gamma_{Nj} | R_*)$$

where m is the number of loci, R_* the relationship being considered, and γ_{ij} the genotype of individual i at locus j (see Expression (2.38)).

As a first example, suppose there are two individuals and data available for between 1 and 20 loci where each locus has two alleles, $p = (1/2, 1/2)$. The probability of correct inference can be estimated as before from the number of replicate trials in which the simulated relationship is ranked first. The proportion of samples correctly inferred from 10,000 trials is shown in Figure 3.5 for each relationship. There is an increase in the number of samples correctly inferred as the number of loci increases, but the increase is not smooth. For the example where there are two alleles at each locus, the probability

of correct inference is greatest when there is an odd number of loci given that the pair are sibs. For an unrelated pair, the probability of correct inference is higher when there is an even number of loci. The same pattern is seen when the number of alleles at each locus is increased to four, with the frequency of each allele equal to $1/4$ (see Figure 3.5).

When there are eight equifrequent alleles at each locus, the probability of correct inference for R_2 increases sharply with increasing number of loci; with data from 10 loci the probability of correct inference is close to one. However, for R_1 , initially the number of samples correctly inferred *decreases* with increasing number of loci, then rises slowly. Data from five loci gives a similar probability of correct inference as with data from one locus, with two, three and four loci giving a lower probability.

The overall probability of correct inference is given by Expression (3.3), i.e.

$$\begin{aligned} \Pr(\text{Correct Inference}) &= \Pr(\text{sampling individuals related as } R_* \text{ and inferring } R_*) \\ &= \sum_{\text{Relationships}} \Pr(\text{inferring } R_* \mid R_*) \Pr(R_*) \end{aligned}$$

Figure 3.5 shows the probability of correct inference when there are two individuals for the case where $\Pr(R_*) = 1$, for $R_* = R_1$ or R_2 . The probability of correct inference does not increase smoothly with increasing numbers of loci. If some prior probability for R_1 and R_2 is assumed, then the overall probability of correct inference can be found from Expression (3.3). Taking the example where each locus has two alleles, Figure 3.5a shows the probability of correct inference when $\Pr(R_1) = 0$ (---), 0.5 (...), 0.75 (. . .) or 1 (—), these for 0 and 1 being the same as in Figure 3.5. Clearly, when $\Pr(R_1) = \Pr(R_2) = 0.5$ the probability of correct inference increases monotonically with increasing numbers of loci. However, if the prior probabilities are not equal the increase in the probability of correct inference is not smooth. If $\Pr(R_1)$ is greater than 0.75 there is an initial decrease when using data from two loci.

3.2.3 A Closer Look at Two Loci

Results from simulation experiments show an unexpected change in the probability of correct inference when data from two (or a few) loci are used. When $\Pr(R_1) = 1$, the probability of correct inference is *lower* with two loci than with one locus, whereas if $\Pr(R_2) = 1$ then the probability increases from one to two loci but falls again when

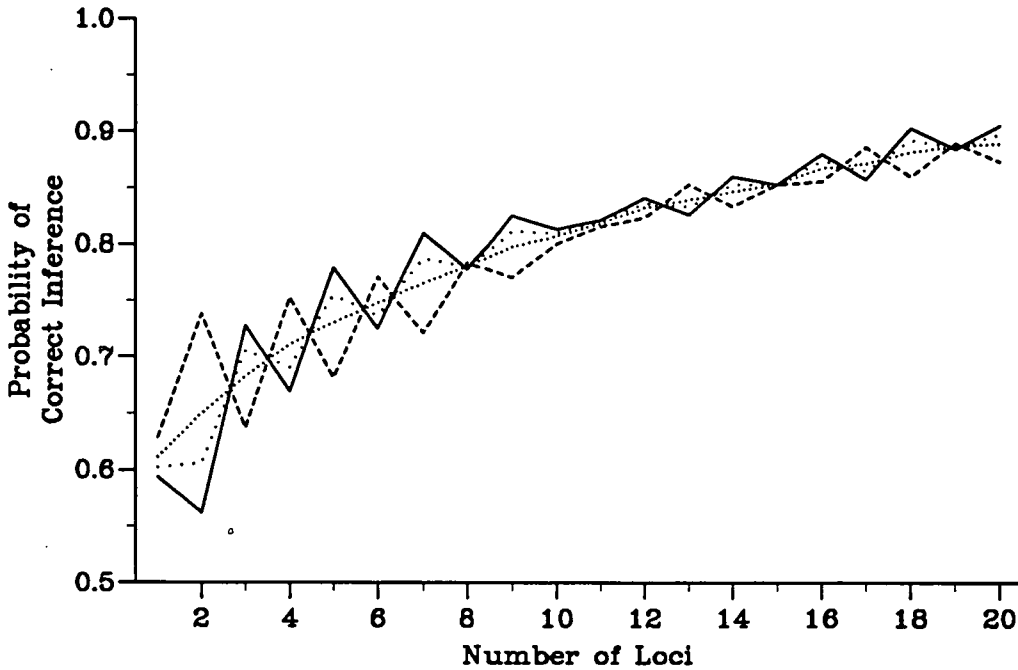


Figure 3.5a: The estimated probability of correct inference from simulation experiments for a sample of two individuals with data from 1 to 20 loci, each locus has two alleles, each with frequency 0.5.

three loci are used (see Figure 3.5). To investigate this further, it is straightforward to calculate all possible pairs and find the expected inference. As before, $\Pr(\Delta_i | R_*)$ is calculated for each locus separately and the two-locus probabilities are simply obtained by multiplying together the appropriate expressions from Table 3.1. The relationship that has the highest likelihood is inferred to be the true one.

For a single locus that has two alleles, there are four possible family patterns, Δ_1 to Δ_4 . Figures 3.1 and 3.2 show the inference of relationship for each pattern; this depends on the gene frequency in some samples. Every two-allele locus, for which there are genotypes available, can be classified as one of these family patterns, and hence for two independent loci there are 16 possible combinations, Δ_1, Δ_1 or Δ_1, Δ_2 etc. It is possible to find the likelihood ratio for each relationship for any combination of gene frequencies for two loci, and hence identify the relationship that will be inferred in each case. Let the frequency of allele A_i be p_i for the first locus and the frequency of allele A_i be q_i for the second locus. For each sample, a likelihood ratio can be calculated in terms of p_i and q_i . If this ratio is greater than one for a particular combination of

frequencies then the two individuals are inferred to be sibs, but if it is less than one then they are inferred to be unrelated (if the ratio is exactly one then no inference can be made). By substituting a range of values for the allele frequency, and evaluating the likelihood ratio, a simple diagram can be drawn to show the inferred relationship. Over a range of allele frequencies, the two loci are indistinguishable and only ten of the 16 possible family pattern combinations need be looked at in detail. Further, of the ten distinct plots only six include a change of inference, the others show a constant inference independent of allele frequency.

1. $\Delta_1\Delta_1$: The two individuals are identical homozygotes at both loci. For locus 1,

$$\Pr(\Delta_1 | R_1) = \frac{1}{4}p_i^2(1 + p_i)^2 \text{ and } \Pr(\Delta_1 | R_2) = p_i^4,$$

and for locus 2,

$$\Pr(\Delta_1 | R_1) = \frac{1}{4}q_i^2(1 + q_i)^2 \text{ and } \Pr(\Delta_1 | R_2) = q_i^4.$$

The overall likelihood ratio is equal to

$$\frac{\Pr(\Delta_1\Delta_1 | R_1)}{\Pr(\Delta_1\Delta_1 | R_2)} = \frac{(1 + p_i)^2(1 + q_i)^2}{16p_i^2q_i^2}. \quad (3.12)$$

When Expression (3.12) is evaluated for a range of gene frequencies it is found that the ratio is always greater than one, and so two individuals with these genotypes will always be inferred to be siblings.

2. $\Delta_1\Delta_2$ or $\Delta_2\Delta_1$: At locus 1 the two individuals are identical homozygotes, and they are both heterozygotes at locus 2. The likelihood ratio

$$\frac{\Pr(\Delta_1\Delta_2 | R_1)}{\Pr(\Delta_1\Delta_2 | R_2)} = \frac{(1 + p_i)^2(q_i - 2q_i^3 + q_i^4)}{16p_i^2q_i^2(1 - q_i)^2} \quad (3.13)$$

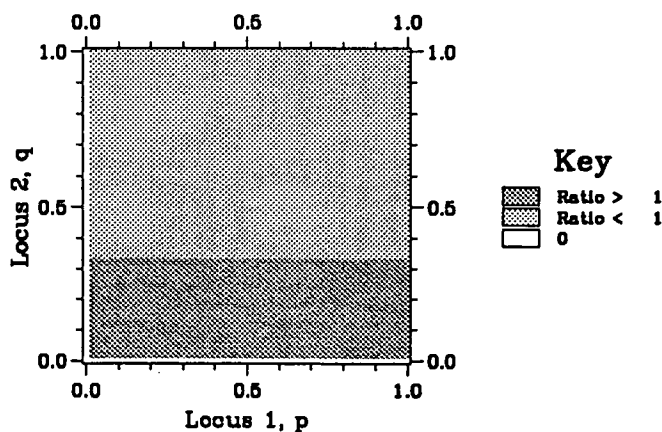
is always greater than one, so these identical individuals are inferred as siblings.

3. $\Delta_1\Delta_3$ or $\Delta_3\Delta_1$: The individuals are identical homozygotes (A_iA_i) for locus 1 but are different homozygotes (A_iA_i, A_jA_j) for locus 2. In this example, the likelihood ratio

$$\frac{\Pr(\Delta_1\Delta_3 | R_1)}{\Pr(\Delta_1\Delta_3 | R_2)} = \frac{(1 + p_i)^2}{16p_i^2} \quad (3.14)$$

is greater than one if the shared allele at locus 1 is less than 1/3, but is less than one if this allele is common. Therefore the inference of relationship depends

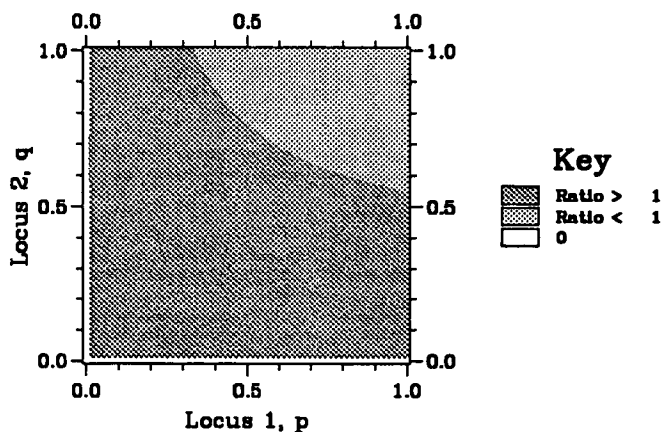
on the gene frequency at one locus. The following diagram illustrates how the inference changes with gene frequency.



4. $\Delta_1\Delta_4$ or $\Delta_4\Delta_1$: The individuals are identical homozygotes (A_iA_i) at locus 1, and have genotypes A_iA_i, A_iA_j at locus 2. The likelihood ratio is

$$\frac{\Pr(\Delta_1\Delta_4 | R_1)}{\Pr(\Delta_1\Delta_4 | R_2)} = \frac{(1 + p_i)^2(1 - q_i^2)}{16p_i^2q_i(1 - q_i)}. \quad (3.15)$$

The inference of relationship depends on the allele frequency at both loci.



5. $\Delta_2\Delta_2$: Both individuals are heterozygotes at both loci. The likelihood ratio is

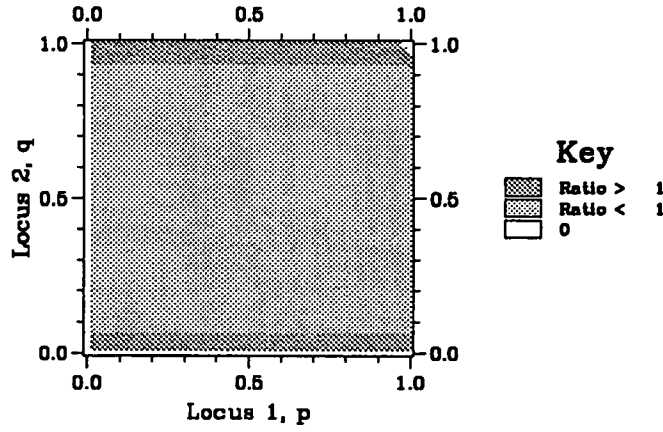
$$\frac{\Pr(\Delta_2\Delta_2 | R_1)}{\Pr(\Delta_2\Delta_2 | R_2)} = \frac{(p_i - 2p_i^3 + p_i^4)(q_i - 2q_i^3 + q_i^4)}{16p_i^2(1 - p_i)^2q_i^2(1 - q_i)^2}. \quad (3.16)$$

As before, for identical individuals the ratio is always greater than 1 and so they are always inferred to be siblings regardless of the allele frequency at either locus.

6. $\Delta_2\Delta_3$ or $\Delta_3\Delta_2$: Here, both individuals are heterozygotes at one locus, but different homozygotes at the other locus. The likelihood ratio is

$$\frac{\Pr(\Delta_2\Delta_3 | R_1)}{\Pr(\Delta_2\Delta_3 | R_2)} = \frac{(p_i - 2p_i^3 + p_i^4)}{16p_i^2(1 - p_i)^2q_i^2}. \quad (3.17)$$

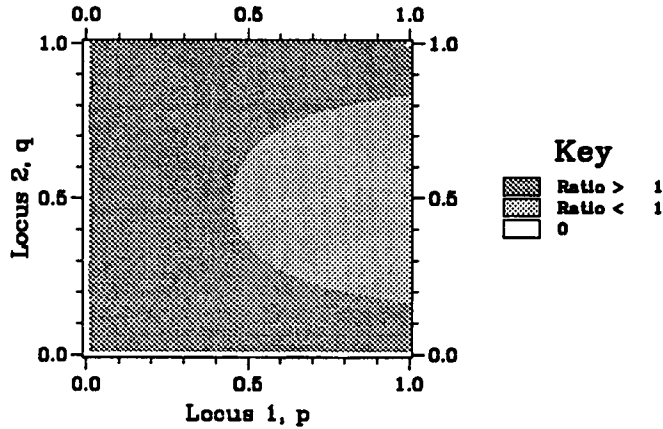
This diagram shows that the pair in the sample are inferred as sibs only if p_i (the allele frequency at the locus for which both are heterozygous) is extreme.



7. $\Delta_2\Delta_4$ or $\Delta_4\Delta_2$: At locus 1 both individuals are heterozygotes, while for locus 2 one is a homozygote and the other is a heterozygote.

$$\frac{\Pr(\Delta_2\Delta_4 | R_1)}{\Pr(\Delta_2\Delta_4 | R_2)} = \frac{(p_i - 2p_i^3 + p_i^4)(1 - q_i^2)}{16p_i^2(1 - p_i)^2q_i(1 - q_i)}. \quad (3.18)$$

This type of sample shows a more complex inference pattern, being dependent on the allele frequency at both loci.



8. $\Delta_3\Delta_3$: The sample here consists of two individuals that are different homozygotes for both loci. The likelihood ratio

$$\frac{\Pr(\Delta_3\Delta_3 | R_1)}{\Pr(\Delta_3\Delta_3 | R_2)} = \frac{1}{16} \quad (3.19)$$

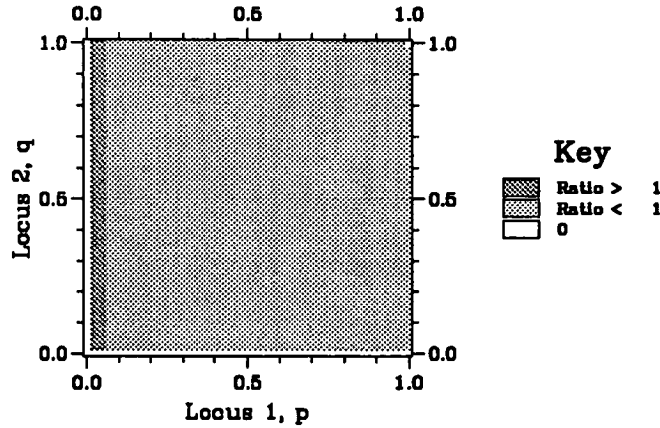
shows clearly that this sample type will always be inferred to be unrelated regardless of the allele frequency at either locus.



9. $\Delta_3\Delta_4$ or $\Delta_4\Delta_3$: For locus 1 the individuals are different homozygotes, and one heterozygote and one homozygote for locus 2. The likelihood ratio is

$$\frac{\Pr(\Delta_3\Delta_4 | R_1)}{\Pr(\Delta_3\Delta_4 | R_2)} = \frac{(1 - q_i^2)}{16q_i(1 - q_i)}. \quad (3.20)$$

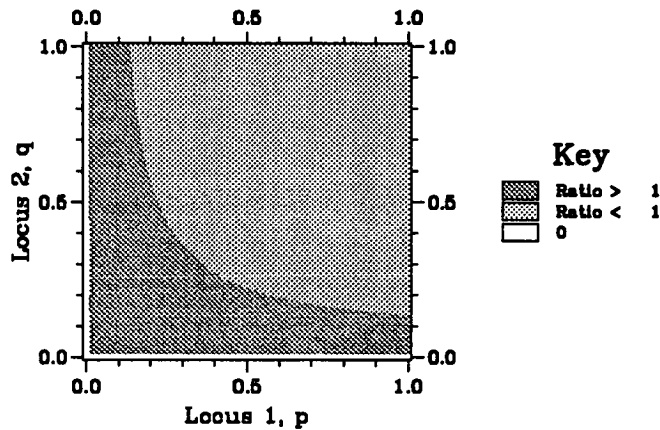
These individuals will be inferred as unrelated except when the shared allele at locus 2 is rare (i.e. A_i for A_iA_i, A_iA_j .)



10. $\Delta_4\Delta_4$: This sample consists of two individuals that are both homozygous at one locus and heterozygous at the other. The likelihood ratio is

$$\frac{\Pr(\Delta_4\Delta_4 | R_1)}{\Pr(\Delta_4\Delta_4 | R_2)} = \frac{(1 - p_i^2)(1 - q_i^2)}{16p_i(1 - p_i)q_i(1 - q_i)}, \quad (3.21)$$

and the inference of relationship depends on the allele frequency at both loci.



The Probability of Correct Inference with Data from Two Loci

The probability of correct inference of relationship for the simple case of two individuals with one two-allele locus is shown in Figure 3.3. Intuitively, it seems reasonable to expect this probability to increase as more information is used but simulation results

show that this is not necessarily the case. The probability of correct inference using two loci is now considered in detail.

The two main features of Figure 3.3 are the position of the “boundaries” where there is a jump in the probability of correct inference (caused by the change of inference for some samples at this gene frequency), and the shape of the curve within each region. For samples with data from two loci, the boundaries are shown in Section 3.2.3, for ten representative examples. There are six examples in which there is a change of inference with changing p_i and q_i . However, there are 81 possible ordered samples of two individuals if data from two loci are considered; each sample corresponds to one of these general graphs.

When all 81 possible samples are included with the appropriate inference of relationship, the pattern of inference boundaries is much more complicated than for a single locus. The probability of correct inference within each region is calculated by summing the appropriate probabilities as described for the one-locus example. The algebraic package REDUCE [Hearn, 1987] was used to find the 162 likelihood expressions required in terms of p_i and q_i . These expressions are evaluated for a range of allele frequencies from zero to $1/2$, to find the the probability of correct inference given R_2 and given R_1 . These probabilities are shown in Figures 3.6 and 3.7 respectively. Each graph is symmetrical about $p_1 = 1/2$ and $q_1 = 1/2$ as there are only two loci, and symmetrical about $p_1 = q_1$. It can clearly be seen that the inference of relationship is complex and depends on the allele frequency at each locus. The probability of correct inference can increase or decrease when more data are included. Note that the “wall” of the plots (where the frequency of one allele is zero for one of the loci) is identical to the graph in Figure 3.3. For example, if the two individuals are unrelated then the probability of correct inference is 0.63 for a single locus with two alleles, when each has frequency $1/2$. This is seen in Figure 3.3 and in Figure 3.6 if p for locus 1 is equal to $1/2$. In Figure 3.6, it is seen that if the second locus has one rare allele, then the probability of correct inference rises to 0.70; if the second locus has one allele with frequency approximately $1/3$, then the proportion of correctly inferred samples drops to 0.52, but rises again to 0.73 if both loci have two alleles with frequency $1/2$. In contrast, if the pair of individuals are full sibs then the probability of correct inference is 0.59 for a single locus with two alleles each with frequency $1/2$ (see Figures 3.3 and 3.7). When

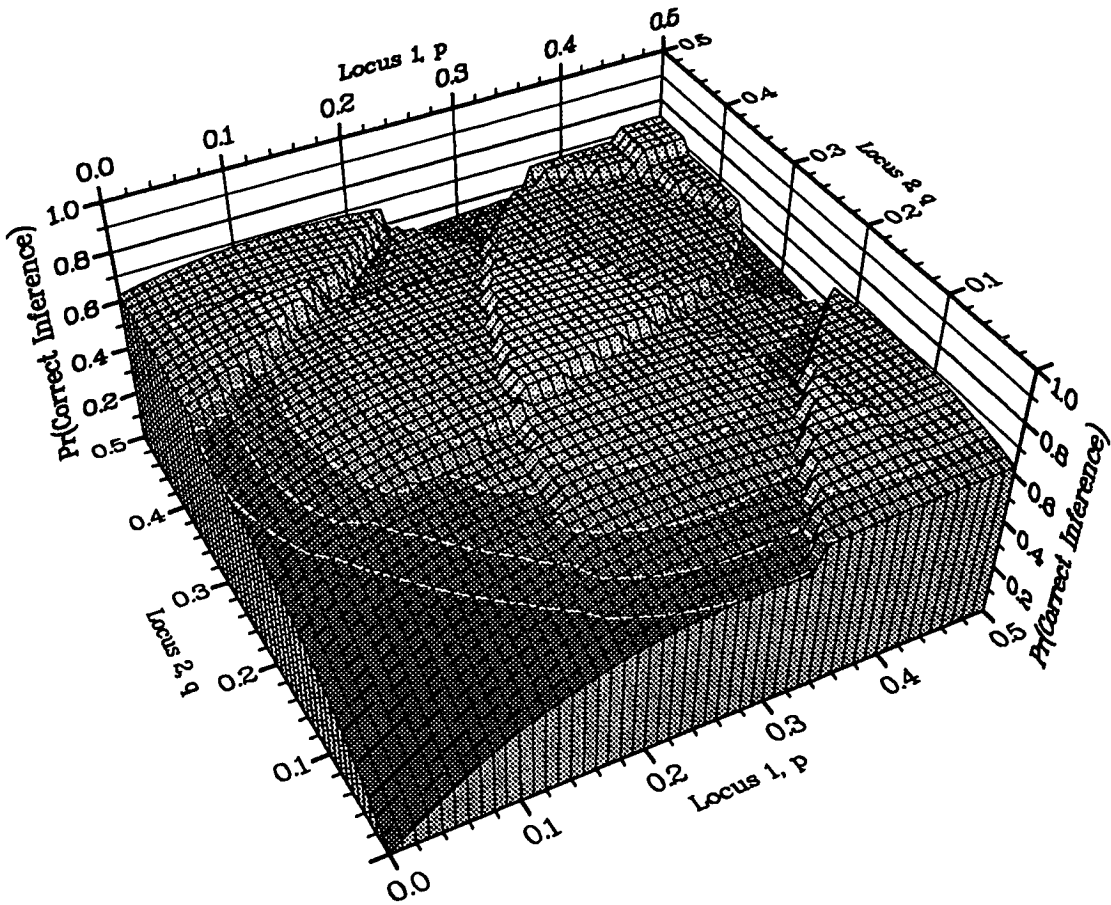


Figure 3.6: The Probability of Correct Inference, given R_2 with data from two two-allele loci.

a second locus is included the proportion of correctly inferred samples decreases very slightly if the second locus has one rare allele, increases to around 0.75 if the second locus has one allele with frequency approximately 0.25, but drops to 0.56 if the second locus has two alleles with equal frequency.

3.2.4 Field Parameters for Two Individuals

Throughout this chapter, simulation experiments for multilocus examples use the same allele frequency distribution for each locus. In practice, each locus would have a different number of alleles, with different frequency distributions. In this final section for a sample of size two, examples are given for inference patterns using data typically available from field experiments.

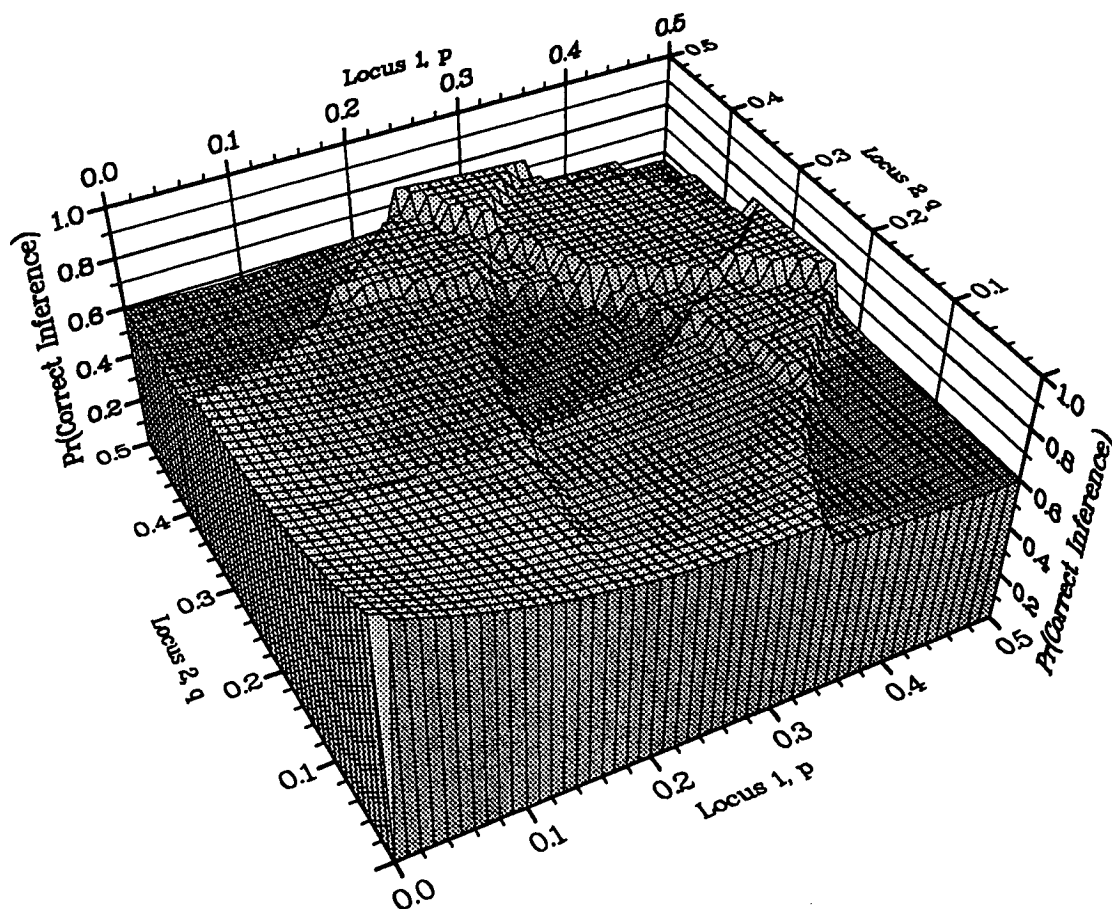


Figure 3.7: The Probability of Correct Inference, given R_1 with data from two two-allele loci.

1. In this first example, the parameters are taken from data supplied by R. H. Thomas (see also Thomas and Barker [1990] and Chapter 6). The allele frequencies for seven loci are estimated from a large population of *Drosophila*.

Locus Name	s	$\underline{p}$
Phosphoglucumutase (<i>Pgm</i>)	3	0.0268, 0.9729, 0.0003
Aldehyde dehydrogenase (<i>Aldox</i>)	2	0.9799, 0.0201
Esterase-1 (<i>Est-1</i>)	3	0.1956, 0.7757, 0.0287
Esterase-2 (<i>Est-2</i>)	4	0.4703, 0.3223, 0.1394, 0.0680
β -N-acetyl-hezosaminidase (<i>Hex</i>)	4	0.9036, 0.0954, 0.0006, 0.0004
Alcohol dehydrogenase (<i>Adh-1</i>)	2	0.6892, 0.3108
<i>Lap</i>	3	0.0380, 0.9368, 0.0252

The simulation programs are used to generate samples of two individuals, either full siblings or unrelated, based on these parameters. The analysis is carried out to find the likelihood ratio and hence the inferred relationship; this is repeated for 10,000 samples. The proportion of samples correctly inferred (with 95% confidence limits) given R_1 is 0.7324 ± 0.0087 and given R_2 is 0.7608 ± 0.0084 .

2. The following parameters are taken from Meagher and Thompson [1987] and are values found for a natural population of a dioecious plant *Chamaelirium luteum* (a member of the lily family).

Locus Name	<i>s</i>	<i>p</i>
Phosphoglucose isomerase	4	0.238, 0.164, 0.559, 0.039
Phosphoglucomutase	3	0.769, 0.072, 0.159
Aspartate aminotransferase ₁	2	0.912, 0.088
Aspartate aminotransferase ₂	3	0.007, 0.992, 0.001
Triose phosphate isomerase ₁	3	0.001, 0.997, 0.002
Triose phosphate isomerase ₂	3	0.003, 0.075, 0.922
Glutamate dehydrogenase	2	0.184, 0.816
Mannose-6-phosphate isomerase	5	0.090, 0.418, 0.481, 0.010, 0.001
Malate dehydrogenase	2	0.426, 0.574
Isocitrate dehydrogenase	3	0.006, 0.789, 0.205
Glyceraldehyde-3-phosphate dehydrogenase	3	0.001, 0.845, 0.154

Samples of two individuals are generated and analysed as before, for 10,000 replicate trials. For this example, the proportion of samples correctly inferred given R_1 is 0.8080 ± 0.0077 and 0.8170 ± 0.0076 given R_2 .

3. In this simulated example, there are five loci with the following allelic frequencies,

Locus	<i>s</i>	<i>p</i>
1	2	0.50, 0.50
2	4	0.50, 0.20, 0.10, 0.20
3	6	0.10, 0.10, 0.30, 0.20, 0.10, 0.20
4	3	0.80, 0.05, 0.15
5	4	0.70, 0.05, 0.05, 0.20

The proportion of samples correctly inferred given R_1 is 0.7791 ± 0.0081 and given R_2 is 0.8252 ± 0.0074 .

Generally, with information from a small number of loci the probability of correct inference depends on the true relationship and on the frequency of the alleles at each locus. As the number of loci increases and more data are incorporated the values

converge; with a sample of just two individuals the probability of correctly inferring the relationship between them is found to be in the range 0.7–0.8.

3.3 A Sample of Three Individuals

When the sample consists of only two individuals, the likelihood for each relationship is always greater than zero (except in the trivial case for $p_i = 0$ or 1). This is true for a sample of any size provided there are only two alleles segregating at every locus of interest. If there are more than two alleles at any one locus, however, it is possible to draw three or more individuals that cannot be full siblings. The probability of obtaining the genotypes observed in the sample is zero and hence the likelihood for the relationship is zero. To demonstrate the effect of this exclusion on the likelihood for each relationship and on the probability of correct inference, the analysis is now extended to a sample of three individuals.

3.3.1 The Likelihood for Each Relationship

When the sample size is increased to three, there are five possible relationships to consider (Section 2.2). If the individuals are labelled X_1, X_2 and X_3 , these five relationships are denoted

1. "All three full sibs"	$R_{1:3}$	as	$\{X_1, X_2, X_3\}$
2. "Two full sibs with X_3 unrelated"	$R_{2:2,1}$	as	$\{X_1, X_2\}\{X_3\}$
3. "Two full sibs with X_2 unrelated"	$R_{2:2,1}$	as	$\{X_1, X_3\}\{X_2\}$
4. "Two full sibs with X_1 unrelated"	$R_{2:2,1}$	as	$\{X_2, X_3\}\{X_1\}$
5. "All three unrelated"	$R_{3:1,1,1}$	as	$\{X_1\}\{X_2\}\{X_3\}$

Note that there are three different genetic relationships described by the notation $R_{2:2,1}$. The likelihoods for each of these five relationships are found as described in Section 2.3.2, and the likelihood ratios are determined by dividing each by the likelihood for $R_{3:1,1,1}$. The five likelihood ratios are ranked in descending order and the relationship in first place is recorded. While two individuals from a family may have genotypes fitting patterns Δ_1 – Δ_4 , Δ_6 , Δ_7 and Δ_{13} only, individuals in a family of three may also fit patterns Δ_5 , Δ_8 , Δ_9 , Δ_{11} and Δ_{12} (the remaining two patterns, Δ_{10} and Δ_{14} require

four or more individuals). For each of these patterns, the probability of obtaining the genotypes given the relationship can be obtained from Expressions (2.5) to (2.32). An important point to note is that some genotype triplets do not fit any of the valid family patterns, for example three different homozygotes. Individuals with these genotypes cannot be full siblings and hence the likelihood for that relationship is zero, so the final ranking in this case would have only four *valid* relationships. No pairs of individuals can be excluded, therefore the likelihood for each of the relationships $R_{2:2,1}$ is always greater than zero.

Consider first the simple case where there are only two alleles segregating at a single locus; there are three possible genotypes and therefore 27 different ordered triplets. These correspond to five of the family patterns (the others require three or more alleles). It is straightforward to calculate all possible likelihoods and hence the inference of relationship made in each case. The simplified expressions are shown in Table 3.2, together with the inference of relationship made for each triplet. This is also shown in Figure 3.8 for a range of allele frequencies and is described below.

1. Pattern Δ_1 : $(A_i A_i, A_i A_i, A_i A_i, i, j = 1, 2; i \neq j)$. When the three individuals are identical homozygotes they are inferred as full sibs for any allele frequency.
2. Pattern Δ_2 : $(A_i A_j, A_i A_j, A_i A_j, i, j = 1, 2; i \neq j)$. If the three individuals are identical they are inferred as full sibs for any allele frequency.
3. Pattern Δ_3 : $(A_i A_i, A_i A_i, A_j A_j, i, j = 1, 2; i \neq j)$. For this pattern, two of the individuals are identical homozygotes and the third is a homozygote for the other allele. For any allele frequency, the relationship $R_{2:2,1}$ is inferred; the two identical individuals are full sibs and the third is unrelated.
4. (a) Pattern $\Delta_4, r_2 = 1$: $(A_i A_i, A_i A_i, A_i A_j, i, j = 1, 2; i \neq j)$. In this type of sample two of the individuals are identical homozygotes and the third is a heterozygote with one allele in common with the other two individuals. If the frequency of the shared allele is low (less than 0.281) then the individuals are inferred as three full sibs. If, however, the shared allele is more common then the relationship $R_{2:2,1}$ is inferred; the heterozygote is unrelated to the two homozygotes.

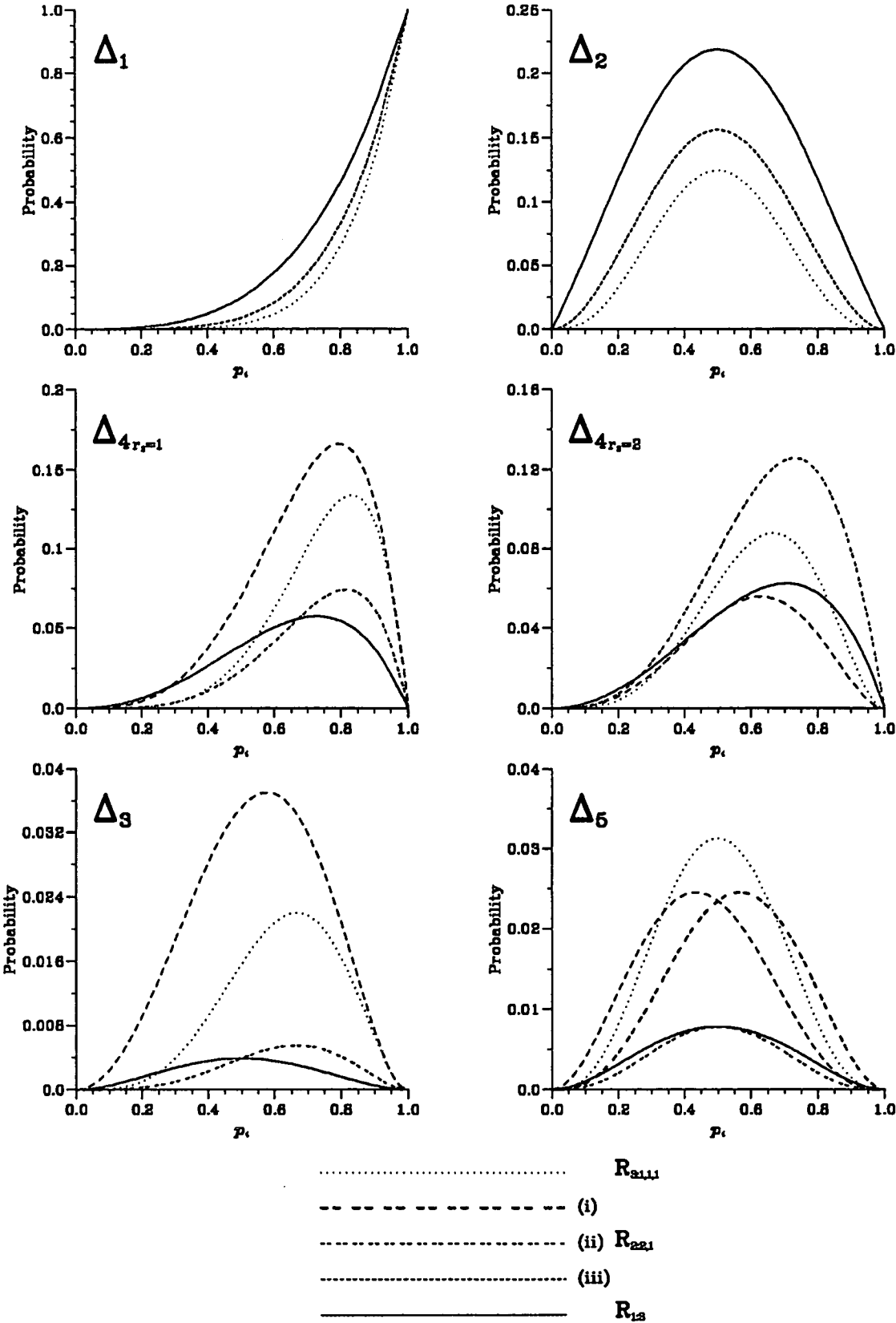


Figure 3.8: The likelihoods for each of the relationships found for a sample of size three, for a single locus with two alleles. The inferred relationship is shown by the curve with the maximum value for a given allele frequency. A full description is given in the text. Note that each graph is drawn to a different scale.

(b) Pattern $\Delta_4, r_2 = 2$: ($A_i A_i, A_i A_j, A_i A_j, i, j = 1, 2; i \neq j$). This pattern is similar to the previous one, except that here there are two heterozygotes and only one homozygote. Here, if the shared allele has frequency less than 0.265, the individuals are inferred as three sibs but if the allele is more common the identical individuals are inferred as a pair of sibs.

5. Pattern Δ_5 : ($A_i A_i, A_i A_j, A_j A_j, i, j = 1, 2; i \neq j$). When the three members of the sample have different genotypes, the inference depends on the allele frequency. If the frequency of allele A_i is less than 1/3 then the $A_i A_i$ homozygote is inferred as a sib of the heterozygote. If the frequency lies between 1/3 and 2/3 then the

Pattern	$\Pr(\Delta_i R_*)$		
	$R_{1,3}$	$R_{2,2,1}$	$R_{3,1,1,1}$
Δ_1	$\frac{1}{16}p_i^2(1+3p_i)^2$	$\frac{1}{4}p_i^4(1+p_i)^2$	p_i^6
R_* inferred:	for $0 < p_i < 1$		
Δ_2	$\frac{1}{2}p_i p_j(1+3p_i p_j)$	$2p_i^2 p_j^2(1+p_i p_j)$	$8p_i^3 p_j^3$
R_* inferred:	for $0 < p_i < 1$		
Δ_3	$\frac{1}{16}p_i^2 p_j^2$	$\{A_i A_i, A_i A_i\}\{A_j A_j\} \uparrow \quad \frac{1}{4}p_i^2 p_j^2(1+p_i)^2$ $\{A_i A_i, A_j A_j\}\{A_i A_i\} \quad \frac{1}{4}p_i^4 p_j^4$	$p_i^4 p_j^2$
R_* inferred:		$\uparrow$ for $0 < p_i < 1$	
$\Delta_4, (r_2 = 1)$	$\frac{1}{8}p_i^2 p_j(1+3p_i)$	$\{A_i A_i, A_i A_i\}\{A_i A_j\} \uparrow \quad \frac{1}{2}p_i^3 p_j(1+p_i)^2$ $\{A_i A_i, A_i A_j\}\{A_i A_i\} \quad \frac{1}{2}p_i^4 p_j(1+p_i)$	$2p_i^5 p_j$
R_* inferred:	if $p_i < 0.281$	$\uparrow$ if $0.281 < p_i < 1$	
$\Delta_4, (r_2 = 2)$	$\frac{1}{4}p_i^2 p_j(1+p_i)$	$\{A_i A_i, A_i A_j\}\{A_i A_j\} \quad p_i^3 p_j^2(1+p_i)$ $\{A_i A_j, A_i A_j\}\{A_i A_i\} \uparrow \quad p_i^3 p_j(1+p_i p_j)$	$2p_i^5 p_j$
R_* inferred:	if $p_i < 0.265$	$\uparrow$ if $0.265 < p_i < 1$	
Δ_5	$\frac{1}{8}p_i^2 p_j^2$	$\{A_i A_i, A_i A_j\}\{A_j A_j\} \uparrow \quad \frac{1}{2}p_i^2 p_j^3(1+p_i)$ $\{A_i A_i, A_j A_j\}\{A_i A_j\} \quad \frac{1}{2}p_i^3 p_j^3$ $\{A_j A_j, A_i A_j\}\{A_i A_i\} \downarrow \quad \frac{1}{2}p_i^3 p_j^2(1+p_j)$	$2p_i^3 p_j^3$
R_* inferred:		$\uparrow$ if $0 < p_i < 1/3, \downarrow$ if $2/3 < p_i < 1$	if $1/3 < p_i < 2/3$

Table 3.2: The likelihood for each relationship for a sample of three individuals, for a single locus with two alleles. The relationship inferred for each pattern is indicated for the appropriate p_i . Note that three individuals will only be inferred as unrelated if they have genotypes matching pattern Δ_5 , ($A_i A_i, A_i A_j, A_j A_j, i, j = 1, 2; i \neq j$) and the frequency of p_i lies between 1/3 and 2/3.

individuals are inferred as unrelated. Note that this is the only pattern in which the individuals can be inferred as unrelated.

3.3.2 The Probability of Correct Inference

The probability of correct inference is calculated in a similar way to that described in Section 3.2.2 for $N = 2$. With two individuals in the sample, there are only two relationships, $R_{1:2}$ and $R_{2:1,1}$ and so either of these may be inferred. When there are more than two individuals, there are of course many more possible relationships that may be inferred. The probability of *correct* inference is found by counting the number of times the simulated (true) relationship is inferred but it is also of interest to find the number of times each of the alternative relationships is inferred.

For a single locus with two alleles it is straightforward to calculate the exact probability of inferring each of the relationships for given experimental set of parameters. The algebraic package REDUCE was used to find theoretical values for both the probability of correct inference and the probability of incorrectly inferring R_* . The results from simulation experiments are in close agreement with the theoretical values.

A Single Locus

The inference of relationship made for each possible relationship is given in Table 3.2. Adding together the appropriate probabilities for given allele frequencies gives the probability of correct inference; this is shown in Table 3.3. The expected probability of correct inference of relationship is shown in Figure 3.9 for a range of p_1 from 0 to 1. At extreme allele frequencies, most of the samples will contain homozygous individuals and these will be inferred to be siblings. Therefore, when this is the simulated relationship, the probability of correct inference is high. If the frequency of one allele is less than $1/3$ then unrelated individuals are never correctly inferred. At intermediate allele frequencies the probability of correct inference is less than one half for all three relationship types. When the three individuals are simulated to be unrelated (randomly selected genotypes), this will not be inferred unless the frequency of one allele lies between $1/3$ and $2/3$; the probability of correct inference is less than 0.2. A marked discontinuity is seen in Figure 3.3 for a sample of two individuals, caused by a change of inference

R	Range of p_i	Probability of correct inference
$R_{1:3}$	$0 < p_1 < 0.2647$	$\frac{1}{4}(3p_1^4 - 18p_1^3 + 27p_1^2 - 12p_1 + 4)$
	$0.2647 \leq p_1 < 0.2808$	$\frac{1}{2}(3p_1^4 - 9p_1^3 + 12p_1^2 - 6p_1 + 2)$
	$0.2808 \leq p_1 < 0.7192$	$\frac{1}{8}(21p_1^4 - 42p_1^3 + 45p_1^2 - 24p_1 + 8)$
	$0.7192 \leq p_1 < 0.7353$	$\frac{1}{2}(3p_1^4 - 3p_1^3 + 3p_1^2 - 3p_1 + 2)$
	$0.7353 \leq p_1 < 1$	$\frac{1}{4}(3p_1^4 + 6p_1^3 - 9p_1^2 + 4)$
$R_{2:2,1}$	$0 < p_1 < 0.2647$	$\frac{1}{4}p_1(12 - 31p_1 + 26p_1^2 - 7p_1^3)$
	$0.2647 \leq p_1 < 0.2808$	$\frac{1}{4}p_1(4p_1^5 - 8p_1^4 - 7p_1^3 + 30p_1^2 - 31p_1 + 12)$
	$0.2808 \leq p_1 < 0.3333$	$\frac{1}{4}p_1(2p_1^5 - 10p_1^4 - 5p_1^3 + 32p_1^2 - 31p_1 + 12)$
	$0.3333 \leq p_1 < 0.6667$	$\frac{1}{4}p_1(6p_1^5 - 18p_1^4 - 5p_1^3 + 40p_1^2 - 35p_1 + 12)$
	$0.6667 \leq p_1 < 0.7192$	$\frac{1}{4}p_1(2p_1^5 - 2p_1^4 - 25p_1^3 + 48p_1^2 - 35p_1 + 12)$
	$0.7192 \leq p_1 < 0.7353$	$\frac{1}{4}p_1(4p_1^5 - 16p_1^4 + 13p_1^3 - 2p_1^2 - 3p_1 + 4)$
	$0.7353 \leq p_1 < 1$	$\frac{1}{4}p_1^2(5 + 2p_1 - 7p_1^2)$
$R_{3:1,1,1}$	$0 < p_1 < 0.3333$	0
	$0.3333 \leq p_1 < 0.6667$	$12p_1^3(1 - p_1)^3$
	$0.6667 \leq p_1 < 1$	0

Table 3.3: The probability of correct inference for a samples of three individuals, using data from one locus with two alleles.

around allele frequency 1/3. A similar discontinuity is seen in Figure 3.9 for a sample of three individuals; there is a switch of inference for some samples when one allele has a frequency 0.26, 0.28 or 0.33.

The observed probabilities of correct inference given R_* from simulation experiments are shown in Table 3.4, each result corresponds to 10,000 trials (shown in **bold type**). Three different relationships are simulated, namely $R_{1:3}$, $R_{3:1,1,1}$ and $R_{2:2,1}$ as $\{X_1, X_2\}\{X_3\}$. The observed results agree well with the values expected (Figure 3.9). In addition to the probability of correct inference, values are also found for the probability of incorrectly inferring R_* . For example, in Table 3.4, when the three individuals are simulated to be unrelated with $p_1 = 0.1$, the probability of correct inference is zero. However, when this is the case, the relationship $R_{1:3}$ will be inferred approximately 53% of the time and each of the three relationships $R_{2:2,1}$ will be inferred approximately 16% of the time. If the frequency of one allele is less than 1/3, $R_{3:1,1,1}$ will never be inferred, either correctly or incorrectly. At more extreme frequencies, most samples will be inferred as full siblings, although the probability of inferring $R_{1:3}$ is greater for

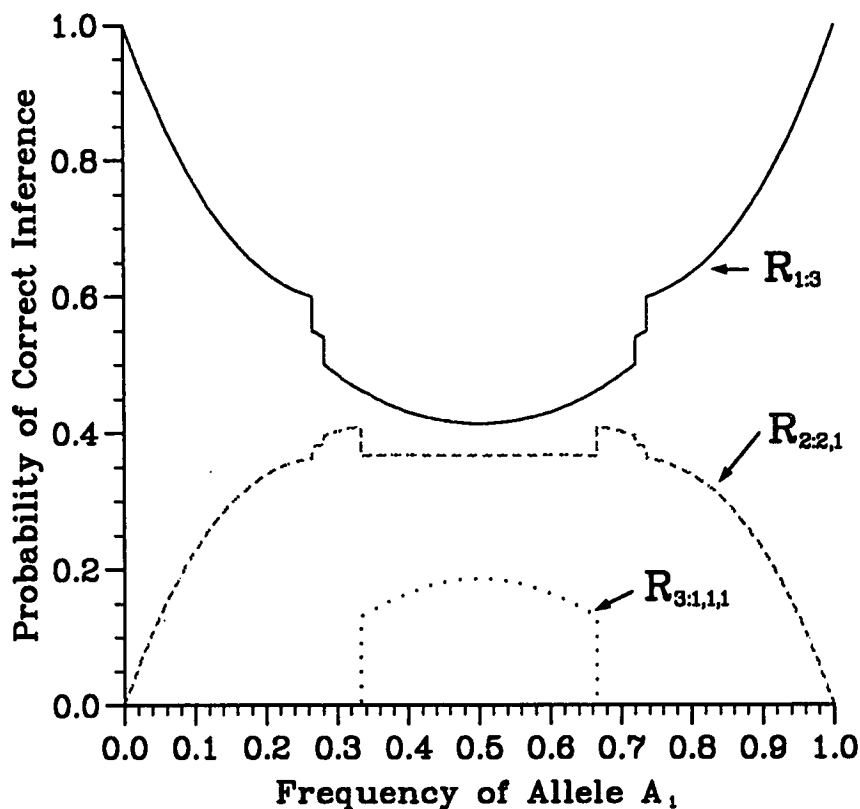


Figure 3.9: The probability of correct inference given each relationship for a sample of three individuals.

more closely related individuals. For the simulation experiments, when $\Pr(R_{1:3}) = 1$ or $\Pr(R_{3:1,1,1}) = 1$ the three relationships $R_{2:2,1}$ are equivalent when taken over repeated replicate trials and therefore give approximately the same probability of correct inference. When $\Pr(R_{2:2,1}) = 1$, there will be two results, depending on whether the true siblings are considered together. Generally, when the individuals are full sibs, if this is not inferred, then they will be inferred as $R_{2:2,1}$. If they are simulated to be $R_{2:2,1}$, then if this is not inferred they will be inferred as either $R_{1:3}$ or an alternative $R_{2:2,1}$. When they are simulated to be unrelated, then the three individuals will usually be inferred as $R_{2:2,1}$.

For loci with more than two alleles, the probability of correct inference is estimated empirically as the proportion of trials in which the simulated relationship is correctly inferred. Examples for a single locus with s alleles are given below, together with examples for multiple, independent loci.

R_*	p_1	Inferred Relationship				
		$R_{1:3}$	$R_{2:2,1}$	$R_{2:2,1}$	$R_{2:2,1}$	$R_{3:1,1,1}$
$R_{1:3}$	0.1	0.7696	0.0791	0.0798	0.0715	0
	0.2	0.6306	0.1247	0.1226	0.1221	0
	0.3	0.4866	0.1717	0.1703	0.1714	0
	0.4	0.4262	0.1828	0.1714	0.1788	0.0408
	0.5	0.4130	0.1829	0.1780	0.1780	0.0481
$R_{2:2,1}$	0.1	0.6104	0.2308	0.0806	0.0782	0
	0.2	0.4186	0.3400	0.1216	0.1198	0
	0.3	0.2828	0.3944	0.1639	0.1589	0
	0.4	0.2365	0.3705	0.1476	0.1509	0.0945
	0.5	0.2254	0.3701	0.1474	0.1487	0.1084
$R_{3:1,1,1}$	0.1	0.5309	0.1565	0.1512	0.1614	0
	0.2	0.3022	0.2322	0.2273	0.2383	0
	0.3	0.1907	0.2665	0.2755	0.2673	0
	0.4	0.1624	0.2159	0.2220	0.2293	0.1704
	0.5	0.1526	0.2165	0.2197	0.2199	0.1913

Table 3.4: The proportion of samples inferred as each of the different relationships given that R_* is the true relationship, for a sample of three individuals, with data from one locus with two alleles. The probability of *correct* inference is shown in bold type.

Multiple Loci

When more data are included in the analysis from several loci, the probability of correct inference will generally be expected to increase. But, as seen for two individuals, it is found that when the individuals are full sibs, a smaller proportion of samples are correctly inferred when two or more loci are included. For example, Table 3.5a shows the proportion inferred as the indicated relationship out of 1000 replicate trials for the case where each locus has two alleles with equal frequency. When relationships $R_{2:1,1}$ or $R_{3:1,1,1}$ are simulated, the probability of correct inference increases as the number of loci increases. When $R_{1:3}$ is simulated, however, there is a decrease from one to two loci but then the probability of correct inference increases again, reaching 0.694 for nine loci. Note that the relationship $R_{3:1,1,1}$ is correctly inferred less than half of the time even with data from nine loci. Table 3.5b shows the proportion of samples inferred as each relationship, for the case where each locus has two alleles each with frequency

R_*	Number of Loci								
	1	2	3	4	5	6	7	8	9
$R_{1:3}$	0.411	0.377	0.483	0.513	0.578	0.587	0.643	0.665	0.694
$R_{2:2,1}$	0.194	0.186	0.208	0.136	0.146	0.118	0.121	0.101	0.114
$R_{3:1,1,1}$	0.045	0.114	0.045	0.057	0.019	0.044	0.015	0.025	0.015
$R_{1:3}$	0.223	0.133	0.174	0.162	0.164	0.147	0.116	0.123	0.120
$R_{2:2,1}$	0.361	0.436	0.534	0.506	0.569	0.562	0.625	0.645	0.680
$R_{2:2,1}$	0.144	0.138	0.142	0.092	0.100	0.077	0.091	0.061	0.059
$R_{3:1,1,1}$	0.109	0.218	0.082	0.184	0.108	0.140	0.107	0.138	0.095
$R_{1:3}$	0.154	0.054	0.081	0.085	0.072	0.065	0.048	0.042	0.045
$R_{2:2,1}$	0.219	0.198	0.251	0.187	0.213	0.168	0.210	0.185	0.172
$R_{3:1,1,1}$	0.175	0.366	0.221	0.385	0.319	0.434	0.332	0.497	0.437

Table 3.5a: The proportion of samples out of 1000 replicates inferred as R_* for an increasing number of loci. Each locus has two alleles with equal frequency. The probability of *correct* inference and the simulated relationship are shown in bold type.

R_*	Number of Loci								
	1	2	3	4	5	6	7	8	9
$R_{1:3}$	0.623	0.420	0.402	0.487	0.561	0.579	0.625	0.641	0.653
$R_{2:2,1}$	0.126	0.194	0.178	0.163	0.145	0.142	0.118	0.112	0.101
$R_{3:1,1,1}$	0	0.026	0.047	0.030	0.044	0.035	0.031	0.027	0.024
$R_{1:3}$	0.428	0.197	0.144	0.172	0.165	0.137	0.232	0.0127	0.108
$R_{2:2,1}$	0.343	0.442	0.460	0.511	0.502	0.555	0.589	0.606	0.633
$R_{2:2,1}$	0.107	0.156	0.130	0.122	0.110	0.092	0.090	0.081	0.065
$R_{3:1,1,1}$	0	0.095	0.157	0.099	0.157	0.146	0.116	0.114	0.133
$R_{1:3}$	0.298	0.094	0.060	0.080	0.063	0.060	0.045	0.031	0.031
$R_{2:2,1}$	0.216	0.234	0.215	0.227	0.187	0.187	0.201	0.179	0.187
$R_{3:1,1,1}$	0	0.189	0.283	0.256	0.374	0.394	0.372	0.454	0.423

Table 3.5b: The proportion of samples from 1000 replicates that are inferred as the relationship indicated, for the case where each locus has two alleles with frequency 0.2 and 0.8. The simulated relationship and the probability of correct inference are shown in bold type.

0.2 and 0.8. With these allele frequencies and one locus, the relationship $R_{3:1,1,1}$ is never inferred regardless of the true relationship in the sample, but when more loci are included the probability of correct inference increases. As the number of loci increases, the probability of correctly inferring a given relationship approaches the same value regardless of the allele frequency, and this value is higher the more closely related the individuals.

If any locus has three or more alleles, then exclusions may occur. In some samples the likelihood for the relationship $R_{1:3}$ will be zero. Tables 3.6a, 3.6b and 3.6c show the estimated probability of inferring relationship R_* given that the true relationship is $R_{1:3}$, $R_{2:2,1}$ and $R_{3:1,1,1}$, respectively. In each case, results are estimates from a set of 1000 replicate trials and each locus (from one to five independent loci) has three alleles.

For example, if the individuals are full sibs, then the probability of correct inference increases as the number of loci increases, except when one allele is very common ($p=(0.9,0.05,0.05)$). Here, the probability of correct inference decreases from 0.747 to 0.483 and increases again to 0.506 with one, four and five loci, respectively. As would be expected, the three $R_{2:2,1}$ relationships show similar results in this case where $R_{1:3}$ is the simulated structure, so only one set is given in Table 3.6a. Note that for extreme frequencies, most samples will consist of identical homozygotes that will be inferred as full sibs – correctly in this instance. When $R_{1:3}$ is the simulated relationship, it cannot be excluded in any replicate. For the other examples, the proportion of trials in which $R_{1:3}$ is excluded is given in parentheses. Exclusion of relationships is considered in detail in Section 4.2.1.

If the individuals are simulated to be two full sibs and the third unrelated, the probability of correct inference increases with increasing data, reaching about 0.68 with five loci (Table 3.6b). The three individuals are excluded as full sibs in up to 25% of samples, depending on the allele frequencies; a higher number are excluded with intermediate allele frequencies over several loci. It is interesting to note that when the correct relationship is not inferred then the individuals are more likely to be inferred as full sibs than the alternative $R_{2:2,1,1}$ (i.e. where the true sibs are split up into different families).

$\underline{p}$	m	$R_{1:3}$	$R_{2:2,1}$	$R_{3:1,1,1}$
0.333333, 0.333333, 0.333334	1	0.284	0.213	0.058
	2	0.540	0.152	0.020
	3	0.610	0.133	0.025
	4	0.687	0.092	0.020
	5	0.713	0.090	0.012
0.1, 0.4, 0.5	1	0.396	0.184	0.059
	2	0.497	0.166	0.027
	3	0.603	0.138	0.029
	4	0.650	0.112	0.018
	5	0.691	0.103	0.020
0.5, 0.3, 0.2	1	0.383	0.187	0.058
	2	0.581	0.147	0.026
	3	0.599	0.140	0.021
	4	0.654	0.111	0.014
	5	0.700	0.092	0.017
0.4, 0.4, 0.2	1	0.354	0.168	0.105
	2	0.585	0.133	0.047
	3	0.607	0.119	0.019
	4	0.694	0.102	0.024
	5	0.699	0.106	0.015
0.9, 0.05, 0.05	1	0.747	0.072	0
	2	0.606	0.122	0.007
	3	0.513	0.168	0.024
	4	0.483	0.142	0.031
	5	0.506	0.172	0.045

Table 3.6a: The proportion of samples that are inferred as each of the relationship structures indicated, for the case where $R_{1:3}$ is the true relationship. Each of the m loci has three alleles with the frequency distribution shown. As the three individuals are simulated to be full sibs, this likelihood for this relationship is always greater than zero.

$\underline{p}$	m	$R_{1:3}$	$R_{2:2,1}$	$R_{2:2,1}$	$R_{3:1,1,1}$
0.333333, 0.333333, 0.333334	1	0.072 (0.041)	0.469	0.158	0.145
	2	0.171 (0.095)	0.494	0.113	0.156
	3	0.138 (0.156)	0.564	0.082	0.141
	4	0.122 (0.180)	0.643	0.064	0.124
	5	0.104 (0.259)	0.678	0.050	0.117
0.1, 0.4, 0.5	1	0.162 (0.028)	0.373	0.138	0.198
	2	0.178 (0.005)	0.501	0.120	0.118
	3	0.153 (0.075)	0.548	0.083	0.138
	4	0.131 (0.101)	0.574	0.075	0.150
	5	0.125 (0.135)	0.617	0.060	0.153
0.5, 0.3, 0.2	1	0.148 (0.045)	0.414	0.143	0.170
	2	0.187 (0.081)	0.486	0.115	0.131
	3	0.144 (0.118)	0.528	0.084	0.164
	4	0.113 (0.153)	0.628	0.069	0.140
	5	0.134 (0.171)	0.662	0.054	0.107
0.4, 0.4, 0.2	1	0.108 (0.046)	0.376	0.119	0.274
	2	0.189 (0.076)	0.470	0.093	0.166
	3	0.164 (0.105)	0.551	0.081	0.133
	4	0.140 (0.159)	0.636	0.077	0.094
	5	0.116 (0.191)	0.673	0.055	0.118
0.9, 0.05, 0.05	1	0.601 (0.001)	0.245	0.058	0.023
	2	0.386 (0)	0.328	0.083	0.090
	3	0.229 (0.004)	0.420	0.104	0.147
	4	0.164 (0.004)	0.451	0.094	0.201
	5	0.127 (0.007)	0.501	0.073	0.215

Table 3.6b: This Table shows the estimated probability of correct inference for relationship $R_{2:2,1}$ as $\{X_1, X_2\}\{X_3\}$ and the proportion of samples that would be inferred as other relationships. Each of the m loci has three alleles with frequencies as shown. The proportion of replicates in which $R_{1:3}$ is excluded is given in parentheses.

$\underline{p}$	m	$R_{1:3}$	$R_{2:2,1}$	$R_{3:1,1,1}$
0.333333, 0.333333, 0.333334	1	0.039 (0.092)	0.280	0.194
	2	0.072 (0.229)	0.200	0.353
	3	0.049 (0.290)	0.192	0.379
	4	0.054 (0.361)	0.148	0.461
	5	0.026 (0.431)	0.130	0.524
0.1, 0.4, 0.5	1	0.090 (0.066)	0.203	0.336
	2	0.087 (0.117)	0.240	0.282
	3	0.065 (0.153)	0.165	0.418
	4	0.049 (0.207)	0.186	0.409
	5	0.049 (0.267)	0.156	0.475
0.5, 0.3, 0.2	1	0.077 (0.081)	0.229	0.261
	2	0.073 (0.163)	0.212	0.309
	3	0.045 (0.239)	0.191	0.377
	4	0.036 (0.302)	0.179	0.470
	5	0.041 (0.358)	0.156	0.459
0.4, 0.4, 0.2	1	0.063 (0.092)	0.196	0.398
	2	0.096 (0.168)	0.161	0.387
	3	0.058 (0.239)	0.176	0.397
	4	0.043 (0.316)	0.154	0.428
	5	0.044 (0.383)	0.166	0.474
0.9, 0.05, 0.05	1	0.486 (0)	0.173	0.041
	2	0.277 (0.004)	0.201	0.168
	3	0.149 (0.004)	0.188	0.301
	4	0.098 (0.010)	0.174	0.417
	5	0.068 (0.011)	0.138	0.524

Table 3.6c: The estimated probability of correct inference for the true relationship $R_{3:1,1,1}$ and the proportion of samples incorrectly inferred as $R_{1:3}$ or $R_{2:2,1,1}$. Each of the m loci has three alleles with the frequencies shown. The proportion of samples in which $R_{1:3}$ is excluded is shown in parentheses.

The probability of correct inference for $R_{3:1,1,1}$ is very low when there is a single locus with three alleles (Table 3.6c). This probability may decrease for two loci, depending on allele frequencies, but increases as more loci are included. This relationship is correctly inferred approximately half the time with five loci if there are three equifrequent alleles or one very common allele. Relationship $R_{1:3}$ is excluded in a higher proportion of samples when the individuals are really unrelated than when they include two sibs, as expected.

3.3.3 Field Parameters for Three Individuals

A more realistic example is one in which the number and frequency of alleles varies at each locus. So, for each of the examples detailed in this section, the three individuals are generated using a number of loci with variable number and distribution of alleles. The number of times that each relationship is inferred out of 10,000 trials is recorded and the probabilities of inference estimated accordingly.

- 1. In this example, the parameters are taken from data supplied by R. H. Thomas [Thomas & Barker, 1990] from experimental data collected from a wild population of *Drosophila*. See Section 3.2.4 for details of the seven loci used.

Simulated Relationship	Proportion inferred as				
	$R_{1:3}$	$R_{2:2,1}$			$R_{3:1,1,1}$
$R_{1:3}$	0.6452	0.1112	0.1115	0.1118	0.0208
$R_{2:2,1}$	0.1794	0.5557	0.0661	0.0671	0.1340
$R_{3:1,1,1}$	0.0912	0.1687	0.1590	0.1631	0.4190

When the three individuals in the sample are drawn from the same family, the probability of correct inference of $R_{1:3}$ is 0.6452. Approximately one third of the time when $R_{1:3}$ is the true relationship, the likelihood for one of the relationships $R_{2:2,1}$ will have the maximum value and be inferred; $R_{3:1,1,1}$ has maximum value only 2% of the time in this case. The sum of the probabilities may be greater than one due to some relationships being jointly ranked first, and thus both are counted in the analysis program. Draws are discussed in Section 4.2.2). If the three individuals are simulated to be two sibs with the third unrelated then this

is correctly inferred 5557 times out of 10,000, with $R_{1:3}$ and $R_{3:1,1,1}$ being inferred 1794 and 1340 times, respectively. The likelihoods for the relationships $R_{2:2,1}$ in which a different pair from the one simulated are considered as sibs rarely have the maximum value. The relationship $R_{3:1,1,1}$ is correctly inferred in less than half of the samples. Approximately half of the samples are inferred as one of the relationships $R_{2:2,1}$, but few are inferred as $R_{1:3}$.

2. For this example, there are four loci with two, three or five alleles with different frequencies. The overall pattern of results is similar to the previous example, with $R_{1:3}$ having the highest probability of correct inference followed by $R_{2:2,1}$ and $R_{3:1,1,1}$.

Locus	s	p
1	3	0.3333, 0.3333, 0.3334
2	3	0.6, 0.3, 0.1
3	2	0.6, 0.4
4	5	0.1, 0.2, 0.3, 0.2, 0.2

Simulated Relationship	Proportion inferred as				
	$R_{1:3}$	$R_{2:2,1}$			$R_{3:1,1,1}$
$R_{1:3}$	0.6447	0.1091	0.1143	0.1162	0.0166
$R_{2:2,1}$	0.1222	0.6177	0.0576	0.0646	0.1389
$R_{3:1,1,1}$	0.0428	0.1619	0.1652	0.1603	0.4705

3. There are five simulated loci in this final example, with between two and seven alleles per locus.

Locus	s	p
1	2	0.3333, 0.6667
2	3	0.05, 0.1, 0.85
3	2	0.95, 0.05
4	7	0.1, 0.3, 0.2, 0.1, 0.05, 0.1, 0.15
5	4	0.01, 0.01, 0.01, 0.97

Simulated Relationship	Proportion inferred as				
	$R_{1:3}$	$R_{2:2,1}$			$R_{3:1,1,1}$
$R_{1:3}$	0.6311	0.1192	0.1200	0.1176	0.0150
$R_{2:2,1}$	0.1207	0.5945	0.0738	0.0731	0.1442
$R_{3:1,1,1}$	0.0391	0.1732	0.1676	0.1756	0.4491

Repeated experiments with various numbers of loci and different numbers and frequencies of alleles give approximately the same results; for $R_{1:3}$ the probability of correct inference is approximately 0.64, for $R_{2:2,1}$ it is approximately 0.59 and for $R_{3:1,1,1}$ it is 0.45. This is lower than the probability of correct inference for the relationships for a sample of two individuals.

3.4 More than Three Individuals

As the sample size increases, the number of possible relationships increases rapidly; the total number of relationships for N is shown in Figure 2.1. In this section, samples of four, five and six individuals are examined to investigate inference patterns and changes in the probability of correct inference as the set of relationships becomes more complex.

Four Individuals

When the sample size is increased to four, there are fifteen relationship structures. The likelihoods for each relationship are calculated using the same procedure as for smaller samples, and the probabilities of inferring each relationship are found from simulations.

If the four individuals are closely related, for example $R_{1:4}$ or $R_{2:3,1}$, then the probability of correct inference shows the familiar drop from one to two loci, but rises again when several loci are included in the analysis. With one locus, $R_{4:1,1,1,1}$ is never correctly inferred when there are two alleles and $\underline{p} = (0.5, 0.5)$ or $\underline{p} = (0.2, 0.8)$. For example, for m loci each with two equiprequent alleles the probability of correct inference (estimated from 1000 simulations) for each group of relationships is as follows:

m	$R_{1:4}$	$R_{2:3,1}$	$R_{2:2,2}$	$R_{3:2,1,1}$	$R_{4:1,1,1,1}$
1	0.351	0.304	0.224	0.107	0
2	0.302	0.266	0.242	0.204	0.106
3	0.403	0.354	0.367	0.222	0.044
4	0.445	0.363	0.315	0.295	0.193

For $\underline{p} = (0.2, 0.8)$ and $R_{4:1,1,1,1}$, the proportion correctly inferred (from 1000 replicates) is 0, 0.006, 0.023 and 0.029 for one, two, three and four loci, respectively. The probability of correct inference for this relationship is found to be low for all examples examined.

Similarly, for loci with four equally frequent alleles the probability of correct inference for each possible relationships is as follows:

m	$R_{1:4}$	$R_{2:3,1}$	$R_{2:2,2}$	$R_{3:2,1,1}$	$R_{4:1,1,1,1}$
1	0.325	0.366	0.324	0.235	0.099
2	0.571	0.463	0.320	0.297	0.285
3	0.678	0.606	0.492	0.364	0.195
4	0.790	0.663	0.523	0.492	0.315

The proportion of samples correctly inferred for one locus is much the same as when there are two alleles, but rising to higher values with several loci. The estimated probability of correct inference for a group of four full sibs is 0.79 with four loci each with four alleles, but only about 0.3 for $R_{4:1,1,1,1}$.

Turning to the *Drosophila* field data example (Section 3.2.4) for a sample of four individuals, simulation experiments give the estimated probabilities shown in Table 3.7. The probability of correct inference is shown in bold type; as the individuals become less closely related, this probability decreases. While 60% of families of four sibs are correctly inferred, only 16% of unrelated sets will be accurately identified.

Table 3.7 also gives the proportion of trials in which each relationship is incorrectly inferred for given R_* . For example, when the relationship $R_{1:4}$ is simulated, this is correctly inferred in over half of the samples. The proportion of samples inferred decreases as the relatedness of the sample decreases; here, the likelihood for $R_{4:1,1,1,1}$ has the maximum value in only 0.06% of trials. When the true relationship is $R_{2:2,2}$ as $\{X_1X_2\}\{X_3X_4\}$ (this particular set of sibs is always the simulated relationship in the program), this is correctly inferred in 40% of trials. The relationships $R_{3:2,1,1}$ as $\{X_1X_2\}\{X_3\}\{X_4\}$ and $\{X_3X_4\}\{X_1\}\{X_2\}$ are each inferred in about 10% of the trials; the likelihoods for these two relationships are higher than the other $R_{3:2,1,1}$ because they include the correct pairs of sibs. The relationship $R_{1:4}$ is inferred about 7% of the time and each $R_{2:3,1}$ about 5% of the time. The remaining relationships $R_{3:2,1,1}$ that do not include a correct pair of sibs and the relationship $R_{4:1,1,1,1}$ are each inferred in approximately 1.5% of trials. The relationship $R_{4:1,1,1,1}$ is inferred in only 16% of trials; a total of 49% will be inferred as $R_{3:2,1,1}$, 10% as $R_{2:2,2}$, 17% as $R_{2:3,1}$ and 1% as $R_{1:4}$.

Inferred Relationship		Simulated Relationship				
		$R_{1:4}$	$R_{2:3,1}$	$R_{2:2,2}$	$R_{3:2,1,1}$	$R_{4:1,1,1,1}$
$R_{1:4}$	$\{X_1X_2X_3X_4\}$	0.5763	0.1074	0.0768	0.0375	0.0130
$R_{2:3,1}$	$\{X_1X_2X_3\}\{X_4\}$	0.0638	0.4876	0.0545	0.1128	0.0568
	$\{X_1X_2X_4\}\{X_3\}$	0.0645	0.0318	0.0556	0.1161	0.0586
	$\{X_1X_3X_4\}\{X_2\}$	0.0616	0.0282	0.0525	0.0257	0.0574
	$\{X_2X_3X_4\}\{X_1\}$	0.0650	0.0307	0.0587	0.0276	0.0557
$R_{2:2,2}$	$\{X_1X_2\}\{X_3X_4\}$	0.0463	0.0346	0.4081	0.1211	0.0338
	$\{X_1X_3\}\{X_2X_4\}$	0.0424	0.0340	0.0085	0.0186	0.0348
	$\{X_1X_4\}\{X_2X_3\}$	0.0410	0.0353	0.0083	0.0174	0.0358
$R_{3:2,1,1}$	$\{X_1X_2\}\{X_3\}\{X_4\}$	0.0068	0.0635	0.1053	0.3020	0.0798
	$\{X_1X_3\}\{X_2\}\{X_4\}$	0.0065	0.0626	0.0138	0.0384	0.0804
	$\{X_1X_4\}\{X_2\}\{X_3\}$	0.0064	0.0065	0.0131	0.0375	0.0853
	$\{X_2X_3\}\{X_1\}\{X_4\}$	0.0068	0.0600	0.0136	0.0322	0.0870
	$\{X_2X_4\}\{X_1\}\{X_3\}$	0.0063	0.0066	0.0141	0.0365	0.0839
	$\{X_3X_4\}\{X_1\}\{X_2\}$	0.0061	0.0058	0.1027	0.0281	0.0760
$R_{4:1,1,1,1}$	$\{X_1\}\{X_2\}\{X_3\}\{X_4\}$	0.0006	0.0069	0.0167	0.0507	0.1632

Table 3.7: The estimated probability of inferring relationship R_* for each simulated relationship, using the *Drosophila* field parameters, for 10,000 replicate trials. (Note that the simulated relationship is listed horizontally and the inferred relationship is listed vertically, in contrast to previous examples).

A similar pattern is seen for other groups; relationships that include subsets of true sibships, or relationships in which a true sibship is a subset, are inferred more often than relationships that combine unrelated individuals. This is shown clearly in Figure 3.10, for the example with five individuals.

Field Parameters for Five or Six Individuals

The pattern of inference that is seen for three and four individuals is also apparent as the sample size increases. Using the same seven loci for illustration, the pattern of inference for a sample of five individuals is shown in Figure 3.10. As before, one specific relationship is simulated, for example $R_{3:3,1,1}$ as $\{X_1X_2X_3\}\{X_4\}\{X_5\}$ (code 17), and the probability of inference estimated for each of the 52 relationships. As there are 52 possible relationships, results are shown graphically as a set of histograms.

The pattern of inference is similar to that seen for samples of four individuals. In each case the specific simulated relationship is inferred most often but the probability of correct inference is inversely proportional to the number of families. For this sample size, the probability of correct inference of the relationship $R_{5:1,1,1,1,1}$ is only 0.0415. Any of the 52 possible relationships may be inferred when the individuals represent a random sample of genotypes from the population. For more closely related individuals, however, Figure 3.10 shows clearly how some relationship structures are inferred more often than others. These structures are either those containing subsets of the true relationship when this consists of many sibs, or relationships more closely related than the one simulated. The same pattern is repeated in other examples using different parameters although these are not shown.

Further interpretation of the likelihoods from this and similar observations are discussed in Chapter 4.

When the same procedure is followed for a sample of six individuals, there are 203 hypothetical relationships, with 11 major groups. The estimated probability of correct inference for each group is as follows:

$R_{1:6}$	0.5009	$R_{3:2,2,2}$	0.1631
$R_{2:5,1}$	0.4309	$R_{4:3,1,1,1}$	0.1101
$R_{2:4,2}$	0.3252	$R_{4:2,2,1,1}$	0.0876
$R_{2:3,3}$	0.2953	$R_{5:2,1,1,1,1}$	0.0326
$R_{3:4,1,1}$	0.2529	$R_{6:1,1,1,1,1,1}$	0.0071
$R_{3:3,2,1}$	0.1904		

The same general pattern of inference is seen as described above, but the probabilities are lower overall. In this example, when the individuals are simulated to be unrelated, the relationship $R_{6:1,1,1,1,1,1}$ is not the most frequently inferred; every relationship is inferred approximately equally.

Key

- | | | | |
|-----------------------------------|---------------------------------------|---------------------------------------|---|
| 1. $\{X_1 X_2 X_3 X_4 X_5\}$ | 14. $\{X_2 X_3 X_5\} \{X_1 X_4\}$ | 27. $\{X_1 X_2\} \{X_3 X_4\} \{X_5\}$ | 40. $\{X_2 X_4\} \{X_3 X_5\} \{X_1\}$ |
| 2. $\{X_1 X_2 X_3 X_4\} \{X_5\}$ | 15. $\{X_2 X_4 X_5\} \{X_1 X_3\}$ | 28. $\{X_1 X_2\} \{X_3 X_5\} \{X_4\}$ | 41. $\{X_2 X_5\} \{X_3 X_4\} \{X_1\}$ |
| 3. $\{X_1 X_2 X_3 X_5\} \{X_4\}$ | 16. $\{X_3 X_4 X_5\} \{X_1 X_2\}$ | 29. $\{X_1 X_2\} \{X_4 X_5\} \{X_3\}$ | 42. $\{X_1 X_2\} \{X_3\} \{X_4\} \{X_5\}$ |
| 4. $\{X_1 X_2 X_4 X_5\} \{X_3\}$ | 17. $\{X_1 X_2 X_3\} \{X_4\} \{X_5\}$ | 30. $\{X_1 X_3\} \{X_2 X_4\} \{X_5\}$ | 43. $\{X_1 X_3\} \{X_2\} \{X_4\} \{X_5\}$ |
| 5. $\{X_1 X_3 X_4 X_5\} \{X_2\}$ | 18. $\{X_1 X_2 X_4\} \{X_3\} \{X_5\}$ | 31. $\{X_1 X_3\} \{X_2 X_5\} \{X_4\}$ | 44. $\{X_1 X_4\} \{X_2\} \{X_3\} \{X_5\}$ |
| 6. $\{X_2 X_3 X_4 X_5\} \{X_1\}$ | 19. $\{X_1 X_2 X_5\} \{X_3\} \{X_4\}$ | 32. $\{X_1 X_3\} \{X_4 X_5\} \{X_2\}$ | 45. $\{X_1 X_5\} \{X_2\} \{X_3\} \{X_4\}$ |
| 7. $\{X_1 X_2 X_3\} \{X_4 X_5\}$ | 20. $\{X_1 X_3 X_4\} \{X_2\} \{X_5\}$ | 33. $\{X_1 X_4\} \{X_2 X_3\} \{X_5\}$ | 46. $\{X_2 X_3\} \{X_1\} \{X_4\} \{X_5\}$ |
| 8. $\{X_1 X_2 X_4\} \{X_3 X_5\}$ | 21. $\{X_1 X_3 X_5\} \{X_2\} \{X_4\}$ | 34. $\{X_1 X_4\} \{X_2 X_5\} \{X_3\}$ | 47. $\{X_2 X_4\} \{X_1\} \{X_3\} \{X_5\}$ |
| 9. $\{X_1 X_2 X_5\} \{X_3 X_4\}$ | 22. $\{X_1 X_4 X_5\} \{X_2\} \{X_3\}$ | 35. $\{X_1 X_4\} \{X_3 X_5\} \{X_2\}$ | 48. $\{X_2 X_5\} \{X_1\} \{X_3\} \{X_4\}$ |
| 10. $\{X_1 X_3 X_4\} \{X_2 X_5\}$ | 23. $\{X_2 X_3 X_4\} \{X_1\} \{X_5\}$ | 36. $\{X_1 X_5\} \{X_2 X_3\} \{X_4\}$ | 49. $\{X_3 X_4\} \{X_1\} \{X_2\} \{X_5\}$ |
| 11. $\{X_1 X_3 X_5\} \{X_2 X_4\}$ | 24. $\{X_2 X_3 X_5\} \{X_1\} \{X_4\}$ | 37. $\{X_1 X_5\} \{X_2 X_4\} \{X_3\}$ | 50. $\{X_3 X_5\} \{X_1\} \{X_2\} \{X_4\}$ |
| 12. $\{X_1 X_4 X_5\} \{X_2 X_3\}$ | 25. $\{X_2 X_4 X_5\} \{X_1\} \{X_3\}$ | 38. $\{X_1 X_5\} \{X_3 X_4\} \{X_2\}$ | 51. $\{X_4 X_5\} \{X_1\} \{X_2\} \{X_3\}$ |
| 13. $\{X_2 X_3 X_4\} \{X_1 X_5\}$ | 26. $\{X_3 X_4 X_5\} \{X_1\} \{X_2\}$ | 39. $\{X_2 X_3\} \{X_4 X_5\} \{X_1\}$ | 52. $\{X_1\} \{X_2\} \{X_3\} \{X_4\} \{X_5\}$ |

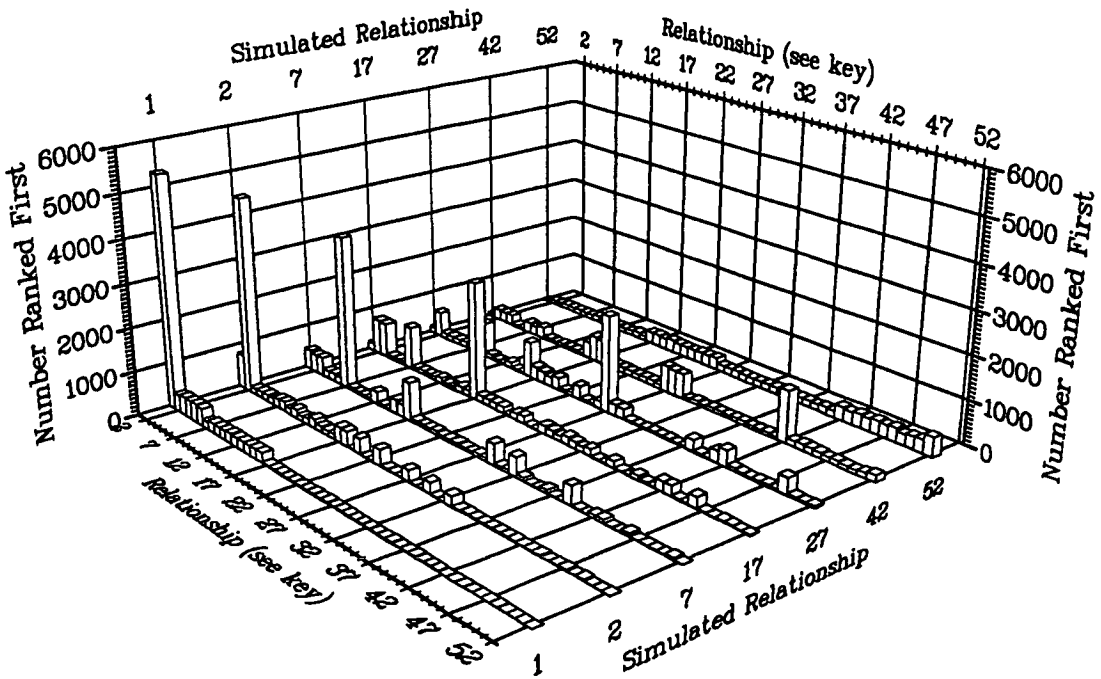


Figure 3.10: The number of trials in which each relationship is inferred, given different simulated relationships for a sample of five individuals.

3.5 Summary

The probability of obtaining the genotypes observed in a sample of N individuals is readily calculated for a given set of genotypes and hypothetical genetic relationship. This is termed the likelihood for the relationship. The hypothesis that has the highest value is an estimate of the relatedness of the individuals in the sample.

If the sample has only two members, then there are only two possible relationships, R_1 and R_2 . It is straightforward to calculate the likelihood for each relationship, and hence find the inference of relationship for any set of genotypes that may occur. Using data from a single locus it is found that two individuals are always inferred as sibs if they have identical genotypes and are always inferred as unrelated if they have no alleles in common, for any allele frequencies. If, however, the individuals have an allele in common, the inference of relationship depends on the frequency of this shared allele; if the frequency is high ($p_i > 1/3$) the individuals are inferred as unrelated, but if it is rare they are inferred as sibs. This pattern of inference is also seen for larger samples. If two or more individuals have identical genotypes for one locus, or if they share a rare allele, they will be inferred as full sibs.

When genotype data from several independent loci are included in the analysis, the overall likelihood ratio for all loci combined is the product of the likelihood ratios for each locus separately. The inference of relationship is based on the overall likelihood ratio and so depends on the genotypes and allele frequency distribution for each locus.

The probability of correctly inferring a particular relationship can be calculated if the prior probability of the relationship is known and the inference of relationship is found for all possible combinations of genotypes.

For a single locus with two alleles, if one allele is rare the probability of correct inference is high for relationships with only one family but very low for relationships with many families. As the frequency of the allele increases, the probability of correct inference decreases for $R_{1:N}$ and increases for $R_{N:1,1,\dots}$. The change is not smooth, however; there are marked discontinuities caused by the switch in inference for some samples around the frequency $p_i = 1/3$. If the two alleles have approximately equal

frequency, the probability of correct inference for $N = 2$ is about 0.6 for both relationships, and is generally lower for larger samples. The number of samples that will be correctly inferred is indirectly affected by the number of alleles, depending more on the relative frequencies of the alleles. For $N = 2$, if there are many alleles with low frequency the probability of correct inference is high for both R_1 and R_2 , but if one allele is common more samples are inferred, correctly or not, as sibs. For larger samples with data from several loci, it is time-consuming to calculate the inference of relationship for every possible combination of genotypes in order to find the probability of correct inference. Computer simulations are therefore used to generate genotypes for individuals with known genetic relationship; the likelihoods are calculated and the inference of relationship found for each simulated sample. The proportion of samples in which the correct simulated relationship is inferred is an estimate of the probability of correct inference. Results from simulation experiments agree well with expected values.

Intuitively, it seems reasonable to expect the probability of correct inference to increase as data from more loci are included. This is shown to be true when there are many loci. The increase is not smooth but rises and falls depending on whether there is an odd or an even number of loci with two alleles. The probability of correct inference is higher if each locus has many alleles, and is close to one for approximately 10 multi-allele loci for $N = 2$. As the sample size increases there are more relationship hypotheses to consider. The pattern of inference is more complex, with switches in inference between different relationships occurring at several allele frequencies. It is found that when genotypes from only one locus are known, the probability of correct inference can be zero for unrelated groups of individuals under certain conditions. When data from several loci are used, some samples may be correctly inferred but the probability of correct inference is low. Groups of unrelated individuals are essentially a random sample of genotypes from the population; it is found that the likelihood for at least one relationship in which the N individuals are members of less than N families will be greater than the likelihood for the relationship $R_{N:1,1,\dots}$.

For sibships greater than two, the likelihood for some relationships can be zero when there are three or more alleles. Hence, some relationship hypotheses can be discounted from the analysis. Exclusion of hypotheses in this way is discussed further in Section 4.2.1.

When there are rare alleles at a locus, the likelihood ratio may be very high for relationship hypotheses that include more than one copy of the rare allele in putative siblings. The total ratio calculated for all loci in these instances will be dominated by the locus with the rare allele, thus increasing the number of samples that will be inferred, rightly or wrongly, as closely related.

As the sample size increases, the probability of correct inference decreases for all relationships. For the majority, this probability is less than one half and is often much lower, revealing that this is not an efficient method for estimating the relatedness of individuals sampled from the same generation.

Taking only the “first passed the post” relationship to estimate the relatedness of the group ignores much information that can be used to infer the degree of relatedness in the sample. In Chapter 4, alternative interpretations of the likelihoods are examined.

Chapter 4

Further Inferences

4.1 Introduction

In Chapter 3 the inference of relationship was based entirely on the relationship hypothesis ranked highest; i.e. the genetic relationship that gives the maximum probability of obtaining the genotypes in the sample. This simplistic view takes no account of the actual value of the likelihoods or the range and magnitudes of the likelihood ratios for the set of relationship structures. After ranking the hypotheses, the order of the relationship structures has been ignored; this ordering may, however, provide further useful information when used in conjunction with the results based on the first rank.

When considering the full set of likelihood ratios, there are two main features that are of interest in providing information for genealogy reconstruction. The first of these is exclusion; the genotypes observed for some samples are incompatible with some relationships. Exclusion of relationships on the basis of the observed genotypes of the individuals may be important for inferring population structure and it is therefore of interest to investigate the proportion of excluded relationships given particular numbers and types of loci. The effect of exclusions on the analysis of larger samples (and the consequences of “draws” in the pattern of ranking) is considered in Section 4.2.

The second feature concerns those relationships not directly excluded; information regarding inference of genetic relationship must be gleaned from the likelihood ratios for valid relationships. If the method is consistent, the true relationship will have the maximum expected likelihood ratio. The expected likelihood ratios for small samples

are investigated in Section 4.3.1 and values obtained from simulation experiments are given in Section 4.3.2. When a relationship structure consists of several families, the likelihood ratios can be simplified; this is shown in Section 4.3.3. In Section 4.4, results from simulation experiments are used to extend the analysis to include all valid relationships in the interpretation of the likelihood ratios.

4.2 The Effect of Exclusions and Draws

4.2.1 Exclusion

Not all the hypothetical relationship structures for a given sample will give a unique, non-zero likelihood. When the likelihood for a relationship is zero, that relationship is said to be excluded. If there are several independent loci, then the overall likelihood for a relationship is the product of the corresponding probabilities for each single locus; exclusion at one locus will obviously exclude the relationship regardless of the genotypes at the other loci (barring mutation or mistyping). The true relationship cannot be excluded: but for other relationships there may be a positive probability of exclusion, if at least one family consists of three or more individuals. Relationships that are not excluded are termed “valid” relationships.

When reconstructing large genealogies extending over many generations, Thompson [e.g. 1974b, 1976] began by finding genotypically compatible parent-pair/offspring triplets. She found that the exclusion of untrue parents was a very important tool in this process; in fact using data from 20 loci, a genealogy of 120 individuals could be determined using exclusion alone. If enough data are available then the only valid pedigree is the correct one.

In order to examine exclusion, Thompson [1976] defines one measure of the “excluding power” of a locus j to be

$$W^{(j)} = -\ln P^{(j)}$$

where $P^{(j)}$ is the probability that randomly chosen unrelated individuals are compatible for a given relationship (in this case, a given parent-pair/offspring triplet). The total

excluding power of m independent loci is given by

$$W = \sum_{j=1}^m W^{(j)}$$

and provides a measure of the information associated with the set of loci with regard to exclusion. If the probability $P^{(j)}$ is high then the triplet of two parents and one offspring are rarely excluded, and hence the power of the locus (with respect to exclusion) is low.

Adapting this to the conditions and notation used in this thesis, a measure of the excluding power of a locus, j , is defined similarly as

$$W^{(j)} = -\ln \Pr(\gamma_{1j}, \gamma_{2j}, \dots, \gamma_{nj}; R_*) \quad (4.1)$$

where $\Pr(\gamma_{1j}, \gamma_{2j}, \dots, \gamma_{nj}; R_*)$ is the probability that n randomly chosen unrelated individuals would have genotypes compatible with relationship structure R_* . This probability is always greater than zero. This measure of excluding power is additive over unlinked loci, so for a set of m such loci the total power with respect to exclusion will be $W = \sum_{j=1}^m W^{(j)}$. Clearly, if $n = 2$, or if there are only two alleles at each locus for any n , the probability that randomly chosen individuals have genotypes compatible with R_* is equal to one and $W^{(j)}$ is zero.

For a single family of n siblings, $W^{(j)}$ is given by $-\ln \Pr(\gamma_{1j}, \gamma_{2j}, \dots, \gamma_{nj}; R_{1:n})$. The probability term is equivalent to

$$\sum_{i=1}^{14} \Pr(\Delta_i | R_{n:1,1,\dots}) \quad (4.2)$$

where Δ_i represents the set of genotypes compatible with family pattern i (see Section 2.3.3). The probability that randomly chosen individuals have genotypes fitting pattern Δ_i are given by Expressions (2.6), (2.8), (2.10), $\dots$, (2.32) for an *ordered* set of genotypes.

So, for any n it is straightforward to find the excluding power of a locus, or set of loci, for the relationship $R_{1:n}$. Table 4.1 shows typical values of $W^{(j)}$ for $n = 2, 3, 4$ or 5 for a selection of single loci. Table 4.2 shows W , the total power with respect to exclusion for two examples with multiple loci. A value of W equal to 0.69 ($= -\ln 0.5$) means that if n individuals are selected at random from a population of unrelated individuals, only half of the samples will be compatible as full siblings. A power, W , equal to 2 means about 86% of such samples would be excluded as siblings. The values in

	<i>s</i>	<i>n</i>				<i>n</i> = 3 and <i>s</i> = 3	
		2	3	4	5	<i>p</i>	<i>W</i>
(a)	2	0	0	0	0	0.025,0.025,0.95	0.000332
	3	0	0.113	0.284	0.489	0.1,0.1,0.8	0.0145
	4	0	0.330	0.888		0.1,0.3,0.6	0.0499
	5	0	0.569	1.491		0.4,0.4,0.2	0.0967
	6	0	0.799			0.3,0.3,0.4	0.109

Table 4.1: (a) Some example values for the excluding power $W^{(j)}$ of a single locus with s equifrequent alleles. This measures the power of a locus to exclude the relationship structure $R_{1:n}$. (b) Further examples for $W^{(j)}$ when $n = 3$, and $s = 3$, with varying $\underline{p}$.

Table 4.1 are exact; they are found by computing $\Pr(\gamma_{1j}, \gamma_{2j}, \dots, \gamma_{nj} \mid R_{n:1,1,\dots})$ for every possible set of genotypes conditional on $\Pr(\gamma_{1j}, \gamma_{2j}, \dots, \gamma_{nj} \mid R_{1:n})$ being greater than zero. The results in Table 4.2 are approximate having been estimated from simulation experiments. For each n , 10,000 samples of n individuals are selected at random from a population of unrelated individuals and if the likelihood for relationship $R_{1:n}$ is found to be greater than zero then a count is incremented. For example, when $n = 4$ for a set of 11 loci, 5764 samples were found to be compatible as sibs whereas 4236 samples were excluded as full sibs. From this, an estimate of $\Pr(\gamma_1, \gamma_2, \dots, \gamma_n, R_{1:n})$ is 0.5764. The estimate of W for these 11 loci for $R_{1:n}$ is therefore $-\ln(0.5764)=0.551$.

The exclusion power of a locus is greatest when the alleles have equal frequency. When the frequency of an allele increases, the frequency of homozygotes increases; thus, more samples will be compatible as sibs on the basis of their genotypes. As the number of alleles at a locus increases, the number of samples containing different heterozygotes will tend to increase; samples with many different heterozygotes will be excluded as sibs. The power of exclusion is therefore higher for loci with many alleles. Results from the examples given in Table 4.2 suggest that the excluding power for a set of loci is not much greater than for a single locus. In both examples, however, several of the loci have only two alleles or several alleles with one very common allele. The excluding power of the individual loci will therefore be zero or very low. As expected, the excluding power of a locus, or set of loci, increases as the sample size increases.

Details of Loci			n			
s	p		3	4	5	6
1	3	0.0268, 0.9729, 0.0003	0.190	0.476	0.821	1.167
2	2	0.9799, 0.0201				
3	3	0.1956, 0.7757, 0.0287				
4	4	0.4703, 0.3223, 0.1394, 0.0680				
5	4	0.9036, 0.0954, 0.0006, 0.0004				
6	2	0.6892, 0.3108				
7	3	0.0380, 0.9368, 0.0252				
1	4	0.238, 0.164, 0.559, 0.039	0.225	0.551	0.872	1.396
2	3	0.769, 0.072, 0.159				
3	2	0.912, 0.088				
4	3	0.007, 0.992, 0.001				
5	3	0.001, 0.997, 0.002				
6	3	0.003, 0.075, 0.922				
7	2	0.184, 0.816				
8	5	0.090, 0.418, 0.481, 0.010, 0.001				
9	2	0.426, 0.574				
10	3	0.006, 0.789, 0.205				
11	3	0.001, 0.845, 0.154				

Table 4.2: The excluding power of two sets of loci (with 7 and 11 loci respectively), for the relationship structure $R_{1:n}$ as obtained from simulation experiments.

For relationship structures that include more than one family, the total excluding power of a locus is equal to the sum of the excluding power for each family separately. For example, if $N = 6$, the excluding power of a locus for the relationship $R_{2:3,3}$ is equal to twice the power of the locus to exclude three individuals as full sibs.

The excluding power of a set of loci is estimated from simulation experiments. When the simulated relationship is $R_{1:N}$, none of the relationship structures can be excluded in the subsequent analysis; if the individuals are compatible as full sibs they are compatible for all other relationship structures considered under the present model. Conversely, if a number of individuals are simulated to be unrelated then the maximum number of relationship structures could be excluded, depending on the exact genotypes observed in each sample. Consider a sample of six individuals; there are 203 possible relationship structures. For clarity, let these relationships be numbered 1–203 as follows:

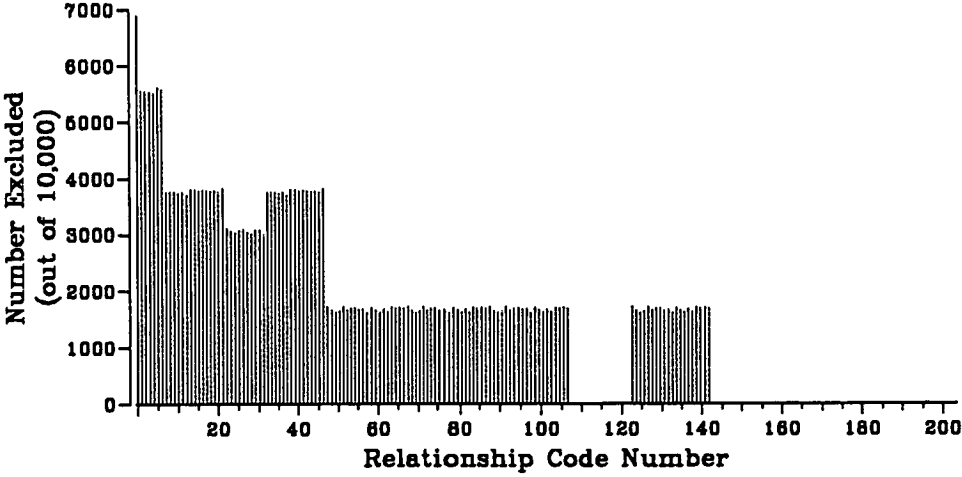


Figure 4.1: The number of samples counted from simulation experiments to be excluded for each relationship structure, given six unrelated individuals.

Codes	Codes	Codes
1 $\Rightarrow R_{1:6}$	33-47 $\Rightarrow R_{3:4,1,1}$	143-187 $\Rightarrow R_{4:2,2,1,1}$
2-7 $\Rightarrow R_{2:5,1}$	48-107 $\Rightarrow R_{3:3,2,1}$	188-202 $\Rightarrow R_{5:2,1,1,1,1}$
8-22 $\Rightarrow R_{2:4,2}$	108-122 $\Rightarrow R_{3:2,2,2}$	203 $\Rightarrow R_{6:1,1,1,1,1,1}$
23-32 $\Rightarrow R_{2:3,3}$	123-142 $\Rightarrow R_{4:3,1,1,1}$	

Figure 4.1 shows the number of samples excluded for each relationship, counted from 10,000 replicate trials, for the case where the true relationship is $R_{6:1,1,1,1,1,1}$ (see Table 4.2 for details of the seven loci parameters used in this experiment). For example, out of 10,000 separate trials, 6888 samples are found to be incompatible as full sibs and thus $R_{1:6}$ is excluded; the estimate for the probability that six unrelated individuals are compatible as sibs ($\Pr(\gamma_1, \gamma_2, \dots, \gamma_6, R_{1:6})$) is therefore 0.3112 and the excluding power of these seven loci is 1.167. Similarly, when the individuals are simulated to be unrelated approximately 5500 samples are excluded for each relationship of the type $R_{2:5,1}$ and the power of exclusion is about 0.8. From the results shown in Figure 4.1, the power of exclusion of these seven loci for each relationship in a sample of six individuals is as follows:

for $R_{1:6}$, $W \approx -\ln 0.3112$	=	1.167	$R_{3:4,1,1} \approx -\ln 0.6237$	=	0.472
$R_{2:5,1} \approx -\ln 0.4453$	=	0.809	$R_{3:3,2,1} \approx -\ln 0.8341$	=	0.181
$R_{2:4,2} \approx -\ln 0.6237$	=	0.472	$R_{4:3,1,1,1} \approx -\ln 0.8341$	=	0.181
$R_{2:3,3} \approx -\ln 0.6957$	=	0.363	else $W = 0$		

Note that relationships coded 108–122 and 143–203 cannot be excluded as they consist of sibships of one or two individuals. The exclusion power for this set of loci is the same for $R_{2:4,2}$ and $R_{3:4,1,1}$, and for $R_{3:3,2,1}$ and $R_{4:3,1,1,1}$, as the small families have no effect on exclusion. The proportion of relationships excluded from the relationship $R_{2:3,3}$ is approximately twice that for $R_{3:3,2,1}$ or $R_{4:3,1,1,1}$ as exclusion depends only on the genotypes in the family of size three.

Using these results for this particular set of loci, it is possible to estimate the proportion of samples that would be excluded for given relationships for any number of individuals. For example, if there were eight individuals in a sample, the excluding power of this set of loci to exclude the relationship $R_{3:4,3,1}$ would be approximately $0.472+0.181+0=0.653$. About half the samples of eight drawn from a population of unrelated individuals would be incompatible for the relationship $R_{3:4,3,1}$ on the basis of the observed genotypes.

If a sample consists of six full sibs, then none of the 203 relationship structures will be excluded. If the six are unrelated, then any of the relationship structures with families of three or more may be excluded. For other relationships, the expected pattern of exclusion is more complex; the number of samples that are excluded will be less than the maximum that is shown in Figure 4.1. The subset of excluded relationship structures reveals information regarding the relationships within the sample. As an example, consider a sample of six individuals simulated to be $R_{2:5,1}$ as $\{X_1, X_2, X_3, X_4, X_5\}\{X_6\}$ (which is allotted the code number 2). Figure 4.2 shows the number of each of the 203 possible relationships that are excluded, again from a total of 10,000 trials. Note that the true relationship (2) is never excluded; in addition, other relationships that include subsets of the true structure are not excluded. Also, if a particular set of three or more individuals are excluded as sibs, then any other relationship structure that includes these sibs as a subset will also be excluded. For single samples, the subset of excluded relationships may give added information on the true relationship. An example is given in Chapter 5.

Thompson [1976] found that a power equal to approximately seven was needed for exclusion to be a useful tool for reconstructing pedigrees from genotypes using individuals from several generations. With this value approximately 0.001 of unrelated pairs are compatible as parents of a given individual, and it is often possible to reconstruct

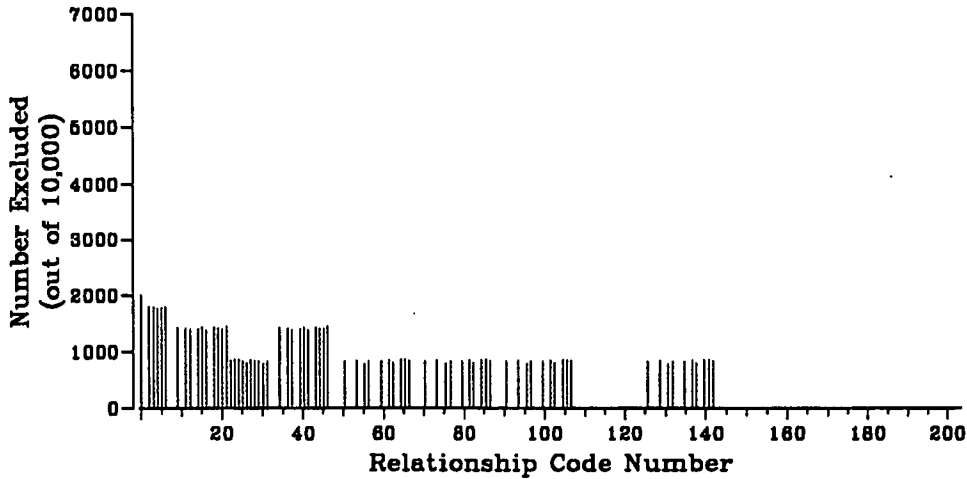


Figure 4.2: The number of samples excluded for each relationship, given that $R_{2:5,1}$ as $\{X_1, X_2, X_3, X_4, X_5\} \setminus \{X_6\}$ is the true relationship. Note that for comparison the same scale is used as in Figure 4.1.

entirely the correct pedigree using exclusion alone. By restricting the analysis to a single generation of individuals, however, much of the importance of exclusion to determine relationships is lost. A large proportion of the set of relationship structures for a given N consist of families of size one or two and these relationships will always have a likelihood greater than zero. Thus, even with a large number of very informative loci, there will always be more than one valid relationship structure and it is not possible to identify the correct relationship based on exclusion alone. If none, or very few, of the relationship structures are excluded for any given sample though, this is evidence that the sample contains full siblings.

4.2.2 Equal Ranking

Occasionally, the likelihoods for two or more relationships for a particular sample of individuals are the same. When this occurs, the relationships will be given equal ranking. This rarely occurs if data are available for more than one locus, but can occur if there is a single locus, especially when one allele is common. When a draw occurs during execution of a simulation run, each relationship is recorded (by the program RANK, Appendix B.1) as having the higher rank. For example, if there are five values 0.01, 0.02, 0.02, 0.03 and 0.06 these will be assigned rank 1, 2, 2, 4 and 5 respectively. As draws tend to occur between relationships in the same group (for

One Locus															
<i>R</i>	Rank														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
<i>R</i> _{1:4}	332	142	223	0	5	0	0	42	207	0	0	49	0	0	0
<i>R</i> _{2:3,1}	92	332	147	0	5	0	0	146	257	0	0	21	0	0	0
	90	332	156	0	5	0	0	131	259	0	0	27	0	0	0
	88	332	136	0	5	0	0	151	261	0	0	27	0	0	0
	79	332	147	0	5	0	0	144	270	0	0	23	0	0	0
<i>R</i> _{2:2,2}	68	14	0	0	0	681	0	0	36	0	13	0	0	188	0
	73	15	0	0	0	681	0	0	32	0	17	0	0	182	0
	87	13	0	0	0	681	0	0	30	0	19	0	0	170	0
<i>R</i> _{3:2,1,1}	13	222	64	0	25	0	0	171	435	0	0	6	64	0	0
	17	213	72	0	21	0	0	168	422	0	0	8	79	0	0
	14	222	73	0	28	0	0	148	428	0	0	12	75	0	0
	18	214	74	0	24	0	0	153	443	0	0	7	67	0	0
	15	211	63	0	27	0	0	163	449	0	0	9	63	0	0
	14	198	73	0	21	0	0	173	436	0	0	7	78	0	0
<i>R</i> _{4:1,1,1,1}	0	49	0	47	207	0	223	0	0	0	0	142	0	0	332
Ten Loci															
<i>R</i>	Rank														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
<i>R</i> _{1:4}	638	156	52	55	32	17	6	11	11	9	4	2	2	5	0
<i>R</i> _{2:3,1}	49	158	130	118	120	109	79	56	54	31	32	19	25	15	5
	56	142	141	127	126	100	78	60	39	38	30	24	24	9	6
	54	143	109	121	112	105	96	77	44	50	27	20	15	13	14
	45	118	159	137	126	111	76	56	52	34	33	18	18	13	4
<i>R</i> _{2:2,2}	36	78	108	87	100	85	107	90	76	66	56	37	20	25	29
	38	55	108	89	89	93	110	87	69	65	59	40	34	28	36
	43	69	104	87	77	93	103	106	67	63	58	39	34	24	33
<i>R</i> _{3:2,1,1}	10	16	28	41	39	47	53	73	90	115	118	119	114	75	62
	4	16	23	23	35	41	67	63	104	103	110	119	121	98	73
	10	24	13	26	52	38	44	69	107	109	107	113	121	106	61
	4	19	20	36	33	48	63	69	97	103	94	135	116	90	73
	4	23	23	26	39	44	62	67	111	99	97	126	112	106	61
	8	14	18	33	39	53	53	68	88	118	113	108	123	99	65
<i>R</i> _{4:1,1,1,1}	2	2	2	9	15	12	17	24	29	37	51	72	86	186	456

Table 4.3: This Table shows the number of times each relationship was assigned to each rank for a total of replicate 1000 trials for a sample of four individuals; the true relationship was $R_{1:4}$ for this example. The data consisted of genotypes known for one or ten loci as indicated; each locus had two equifrequent alleles. When data are available for only one locus, the likelihoods for some relationships will be equal and hence the relationships will be placed in the same rank. This in turn results in “empty” ranks.

example, between all relationships denoted $R_{2,3,1}$), the ranks at which draws are likely to occur are predictable. The effect of draws on the pattern of results is seen clearly in Table 4.3. With a single locus (with two alleles) many draws occur but this is not seen when ten loci are analysed. The presence of relationships with equal ranking suggests that the sample contains several very similar or even identical individuals.

However, the problem of relationships having equal ranking does not generally affect the first rank and the conclusions reached in Chapter 3 are not affected by this. In this chapter, results and conclusions are only influenced when there are few loci or loci with extreme allele frequencies such that many individuals are identical. An alternative method of recording draws would be to assign to each relationship the mean of the ranks that they jointly occupy.

4.3 The Likelihoods and Likelihood Ratios for Valid Relationships

For any sample size, there will always be more than one relationship structure that cannot be excluded, and the relatedness in the sample must be deduced from the likelihoods for the set of non-excluded, or valid, relationships. The actual number of valid relationships depends on the type of loci analysed and on the genotypes observed for the sample. A method to assess information provided by a locus, or set of loci, with respect to exclusion was described in Section 4.2. In this section, the information provided by similar loci for the valid relationships is investigated.

The probability of obtaining the observed genotypes in the sample depends of course on both the number and frequency of alleles at the loci of interest; these probabilities can range from zero to one. Typical values for a sample of two or three individuals are shown in Figures 3.1 and 3.8 respectively; in each example, data are available from a single locus with two alleles. For these examples, the likelihood for each relationship typically lies in the range 10^{-1} to 10^{-2} . As the number of loci increases, the overall probability decreases, reflecting the potentially very large number of combinations of genotypes. The numerical value of the likelihood for a relationship may be very small, particularly if it includes a large sibship, if there are many loci or if there are very rare

alleles at some loci. For example, with ten loci and a sample of two individuals, typical values lie in the range 10^{-10} to 10^{-13} , and for a sample of five individuals with genotype data known for 14 loci the probability of observing a particular set of genotypes may be as small as 10^{-26} to 10^{-31} . An important point to note here concerns the consequence of carrying out real number computation using such small values. It is possible that errors introduced during the calculation of the likelihoods are important when there are many loci and the values are very small.

Although the likelihood for each of the relationships may be very small, it is their values relative to each other that are important. Different hypotheses are therefore compared by calculating a likelihood ratio. The likelihood for each relationship is divided by the likelihood for the relationship $R_{N:1,1,\dots}$ in which all individuals are considered unrelated (see Section 2.4). A consequence of this procedure is discussed further in Section 4.3.3. The natural logarithm of this ratio (or the equivalent log likelihood differences) is used to compare the different hypotheses, and is similar to the support differences used by Edwards [1972] and Thompson [1975,1976].

Note that, throughout, the “likelihood ratio for relationship R_* ” is defined as

$$\frac{\Pr(\text{data} \mid R_*)}{\Pr(\text{data} \mid R_{N:1,1,\dots})},$$

and the log likelihood ratio is the natural logarithm of this expression.

4.3.1 Expected Likelihood Ratios

The mean likelihood ratio (conditional on non-exclusion) for each valid relationship will depend on the true relationship between the individuals. The true relationship should have the maximum expected likelihood ratio if the method is consistent [Edwards, 1972]. In order to assess the reliability of the analysis for inferring genetic relationships, in this section the expected likelihood ratios are found for small sample sizes.

For Two Individuals

Consider again the simple case of a sample consisting of only two individuals. Recall that there are 14 possible genotype patterns, defined by the number and type of

homozygotes and heterozygotes in the sample; two individuals may fit patterns Δ_1 – $\Delta_4, \Delta_6, \Delta_7$ and Δ_{13} (see Section 2.3.3 and Table 3.1).

The likelihood ratio for each possible genotype pattern has been calculated previously (Table 3.2.1), and it is straightforward to find the frequency with which each pattern occurs. Hence, a value for the expected likelihood ratio can be computed.

For any likelihood ratio, it is known that

$$\begin{aligned} E(L(H_1 : H_0) | H_1) &= \int \frac{\Pr(Y | H_1)}{\Pr(Y | H_0)} p(Y | H_1) dY \geq 1 \text{ and} \\ E(L(H_1 : H_0) | H_0) &= \int \frac{\Pr(Y | H_1)}{\Pr(Y | H_0)} p(Y | H_0) dY = 1 \end{aligned}$$

where Y is the data and $p(Y | H_1)$ the frequency of this data under each hypothesis, H_1 or H_0 [Edwards, 1972; Thompson, 1986]. This is confirmed here for this example.

For the simple case of two individuals, the seven possible sample types were given in Table 3.1 and the likelihood ratios listed in Section 3.2.1. If the sample of two individuals are full sibs, the frequency with which genotypes occur that fit each family pattern is given by $\Pr(\Delta_i | R_{1:2})$. The expected likelihood ratio given $R_{1:2}$ is therefore

$$\begin{aligned} \sum_i \frac{\Pr(\Delta_i | R_{1:2})}{\Pr(\Delta_i | R_{2:1,1})} \Pr(\Delta_i | R_{1:2}) &= \sum_{i=1}^s \left(\frac{(1 + p_i)^2}{4p_i^2} \right) \left(\frac{p_i^2(1 + p_i)^2}{4} \right) \\ &+ \sum_{i=1}^s \sum_{\substack{j=2 \\ j < i}}^s \left(\frac{2p_i p_j + p_i + p_j + 1}{8p_i p_j} \right) \left(\frac{p_i p_j (2p_i p_j + p_i + p_j + 1)}{2} \right) \\ &+ 2 \sum_{i=1}^s \sum_{\substack{j=1 \\ j < i}}^s \left(\frac{1}{4} \right) \left(\frac{p_i^2 p_j^2}{4} \right) + 2 \sum_{j=1}^s \sum_{\substack{j=1 \\ j \neq i}}^s \left(\frac{1 + p_i}{4p_i} \right) \left(\frac{p_i^2 p_j (1 + p_i)}{2} \right) \\ &+ \sum_{i=1}^s \sum_{\substack{j=1 \\ k \neq j \neq i}}^s \sum_{k=1}^s \left(\frac{1}{4} \right) \left(\frac{p_i^2 p_j p_k}{2} \right) + \sum_{i=1}^s \sum_{\substack{j=1 \\ k \neq j \neq i}}^s \sum_{k=1}^s \left(\frac{2p_i + 1}{8p_i} \right) \left(\frac{p_i p_j p_k (2p_i + 1)}{2} \right) \\ &+ 2 \sum_{i=1}^s \sum_{\substack{j=1 \\ i < k, j < l, l \neq k \neq j \neq i}}^s \sum_{k=1}^s \sum_{l=1}^s \left(\frac{1}{8} \right) \left(\frac{p_i p_j p_k p_l}{2} \right) \end{aligned} \quad (4.3)$$

where s is the number of alleles at the locus.

Evaluating Expression (4.3) for the simple case of a single locus with two alleles gives an expected likelihood ratio equal to 21/16; for three alleles the expected likelihood ratio is equal to 27/16. The expected ratios are greater than one and independent of allele frequency.

The expression for the expected likelihood ratio can be simplified by letting the s alleles have equal frequency such that $p_i = p = 1/s$ for $i = 1, 2, \dots, s$. Expression (4.3)

then reduces to

$$\frac{s}{16} \left[(1+p)^4 + \frac{1}{2}(s-1)(2p^2+2p+1)^2 + (s-1)p^4 + 4(s-1)p^2(1+p)^2 \right. \\ \left. + 2(s-1)(s-2)p^4 + (s-1)(s-2)p^2(2p+1)^2 + \frac{1}{4}(s-1)(s-2)(s-3)p^4 \right].$$

Substituting $p = 1/s$ gives

$$\frac{2s^5 + 14s^4 + 45s^3 + 18s^2 - 33s + 18}{64s^3}$$

which is equal to 21/16 for $s = 2$, 27/16 for $s = 3$, 4343/2048 for $s = 4$ and 327/125 for $s = 5$, increasing to approximately 11 for 15 equifrequent alleles.

Now, if the pair of individuals are unrelated then a sample of type Δ_i occurs with frequency $\Pr(\Delta_i | R_{2:1,1})$ and the expected likelihood ratio given $R_{2:1,1}$ is

$$\sum_i \frac{\Pr(\Delta_i | R_{1:2})}{\Pr(\Delta_i | R_{2:1,1})} \Pr(\Delta_i | R_{2:1,1}) = \sum_i \Pr(\Delta_i | R_{1:2}) = 1. \quad (4.4)$$

So for a sample of two individuals and a single locus, if $E[L(R_{1:2} : R_{2:1,1}) | R_*]$ denotes the expectation of the likelihood ratio given R_* , then if the pair are full sibs

$$E[L(R_{1:2} : R_{2:1,1}) | R_{1:2}] > E[L(R_{2:1,1} : R_{2:1,1}) | R_{1:2}] = 1.$$

If on the other hand the two individuals are not sibs then

$$E[L(R_{1:2} : R_{2:1,1}) | R_{2:1,1}] = E[L(R_{2:1,1} : R_{2:1,1}) | R_{2:1,1}] = 1$$

and thus all relationships have the same expected likelihood ratio.

When there are multiple loci the expected ratio is equal to the product of the expected value for each independent locus. The *expected* ratio given $R_{2:1,1}$ is always equal to one.

For Three Individuals

The expected likelihood ratios can be found for three individuals in a similar way as for a pair. The notation $E[L(R_* : R_{3:1,1,1})]$ is used in this section to denote the expected value of the ratio

$$\frac{\Pr(\Gamma_1, \Gamma_2, \Gamma_3 | R_*)}{\Pr(\Gamma_1, \Gamma_2, \Gamma_3 | R_{3:1,1,1})}$$

where Γ_i is the genotype for individual i for all loci of interest. Using data from a single locus with two alleles, it is straightforward to calculate the expected likelihood ratios for all relationships and find which has maximum value. The likelihoods for each relationship are given in Section 3.3.1. The following expressions were found using the algebraic package REDUCE.

- Assume that the three sampled individuals are full sibs, $R_{1:3}$.

- ◊ For this example, the expected likelihood ratio for $R_{1:3}$ is

$$E[L(R_{1:3} : R_{3:1,1,1})] = \frac{432p_1^4 - 864p_1^3 + 408p_1^2 + 24p_1 + 1}{256p_1^2(p_1^2 - 2p_1 + 1)}$$

This has a minimum at $p_1 = 1/2$; substituting values for the frequency of p_1 gives, for example, a ratio of 51.013 for $p_1 = 1/100$, 3.211 for $p_1 = 1/10$, 2.223 for $p_1 = 3/10$ and 2.125 for $p_1 = 1/2$.

- ◊ $E[L(R_{2:2,1} : R_{3:1,1,1})]$ given $R_{1:3} = 21/16$

- ◊ $E[L(R_{3:1,1,1} : R_{3:1,1,1})] = 1$

Thus, in this case, the expected likelihood ratio is maximal for the true relationship.

- If the true relationship is $R_{2:2,1}$ (as $\{X_1, X_2\}\{X_3\}$ say), then

- ◊ $E[L(R_{1:3} : R_{3:1,1,1})] = 21/16$,

- ◊ $E[L(R_{2:2,1} \text{ as } \{X_1, X_2\}\{X_3\} : R_{3:1,1,1})] = 21/16$,

- ◊ $E[L(R_{2:2,1} \text{ otherwise} : R_{3:1,1,1})] = 1$ and

- ◊ $E[L(R_{3:1,1,1} : R_{3:1,1,1})] = 1$

The expected ratio for the true relationship is not the maximum.

- Similarly, if the true relationship is $R_{3:1,1,1}$, then all the expected ratios will be equal to one and the true relationship cannot have the maximum expected likelihood ratio.

These expected values are calculated for a single locus with two alleles. If the locus has more than two alleles, then exclusions may occur for the relationship $R_{1:3}$ and the expected likelihood ratio must be calculated conditional on non-exclusion.

4.3.2 Observed Likelihoods and Likelihood Ratios

For small samples of two or three individuals and a single locus, it has been shown that the true relationship does not necessarily have the highest expected likelihood ratio. For larger sample sizes and multiple loci, calculation of the expected ratios is not straightforward. Figures 4.3a to 4.5b show the observed likelihoods and likelihood ratios found from simulation experiments; each set of histograms represents results from 1,000 replicate trials for samples of four individuals. Three different simulated relationships are shown for illustration. Note that when the simulated relationship is any other than $R_{1,4}$, exclusions may occur and the total number of relationships scored is less than the maximum 15 for each of the trials.

When the four individuals are simulated to be full sibs, the log likelihoods for each relationship structure lie between -5 and -45 (Figure 4.3a), with the majority in the range -15 to -20. The likelihood ratios for each relationship are generally greater than one (a log likelihood ratio equal to zero). The ratios are higher for those structures that consider the samples as more closely related, with the true relationship tending to have greater observed likelihood ratios (Figure 4.3b). The log likelihood ratio for $R_{4:1,1,1,1}$ is, of course, always equal to zero.

If $R_{2:2,2}$ is the simulated relationship, a similar distribution of likelihoods is seen although here many of the relationships will be excluded and therefore not included in the graphs (Figure 4.4a). The log likelihood ratios show that the true relationship generally has the higher ratio (Figure 4.4b). In this example, there are two types of relationship $R_{2:2,2}$ depending of the pairing of individuals as sibs. The *incorrect* $R_{2:2,2}$ has a much lower likelihood relative to the other relationships, while the *correct* $R_{2:2,2}$ has higher values.

Finally, Figures 4.5a and 4.5b show the likelihood for each relationship, and the likelihood ratios, for the case where the four individuals are simulated to be not closely related ($R_{4:1,1,1,1}$). The likelihood ratios are predominantly less than one, although the observed values are centred close to one for those relationship structures that consider the individuals as less closely related. In this example, many of the relationships will be excluded.

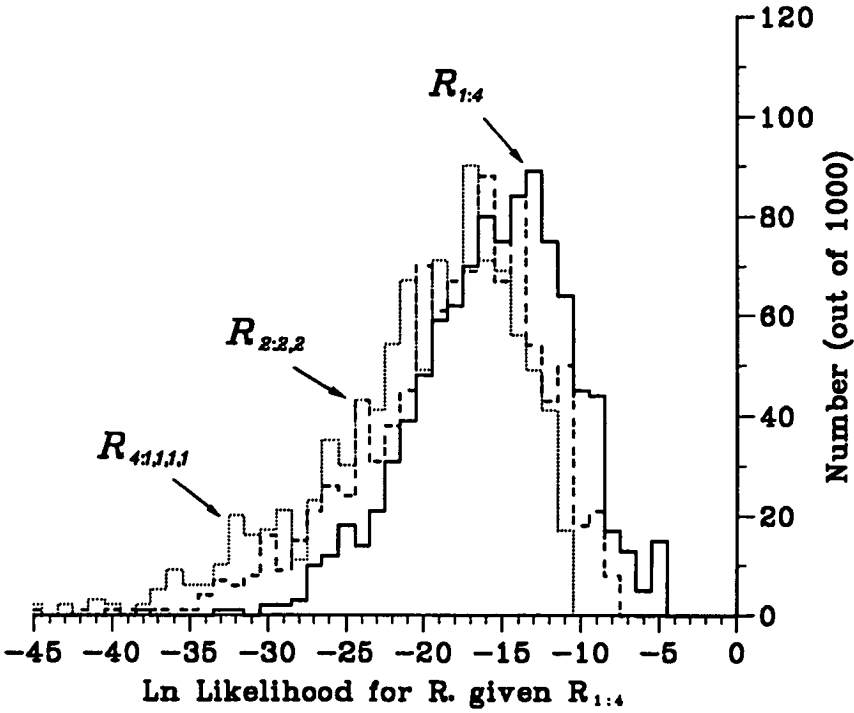


Figure 4.3a: The log likelihood for each relationship observed from simulation experiments for a sample of four individuals, given that $R_{1:4}$ is the true relationship.

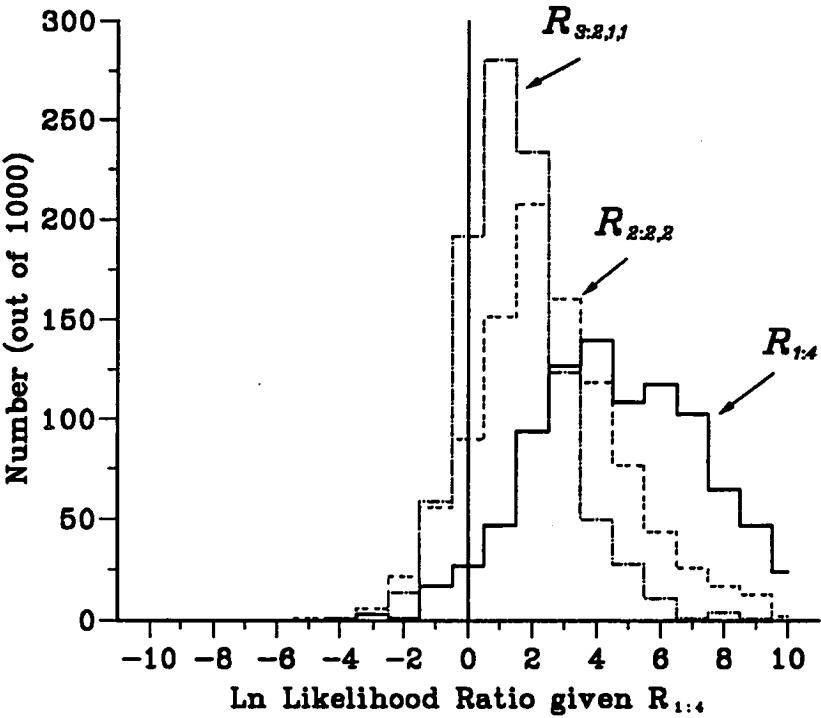


Figure 4.3b: The log likelihood ratios for each relationship, as above. (As expected, the plot for relationship $R_{2:3,1}$ lies between those for $R_{1:4}$ and $R_{2:2,2}$, and the plot for $R_{3:2,1,1}$ in Figure 4.3a lies between $R_{2:2,2}$ and $R_{4:1,1,1,1}$ but these are omitted for clarity.)

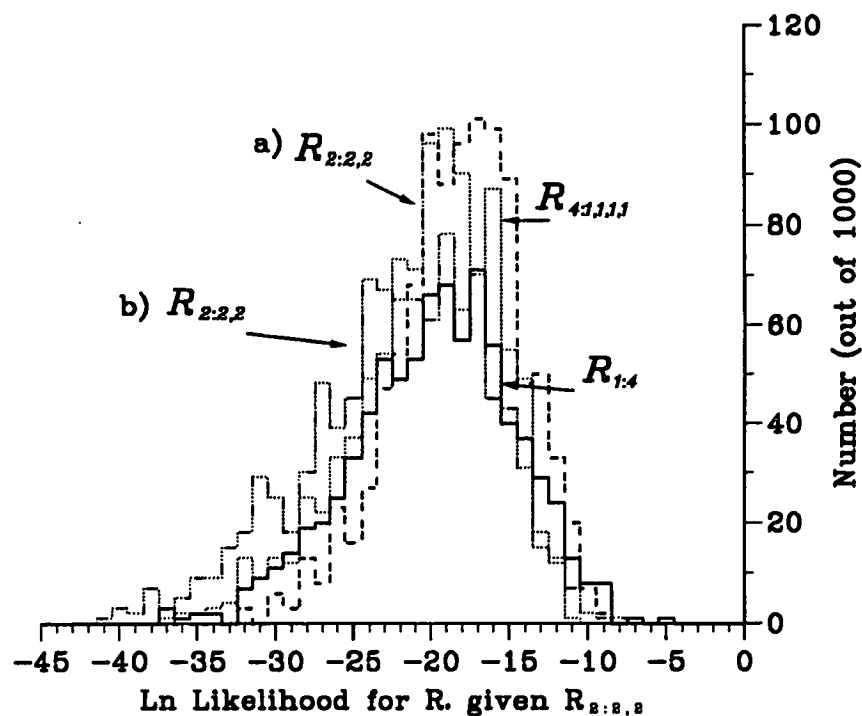


Figure 4.4a: The log likelihood for each relationship observed from simulation experiments for a sample of four individuals, given that $R_{2:2,2}$ is the true relationship.

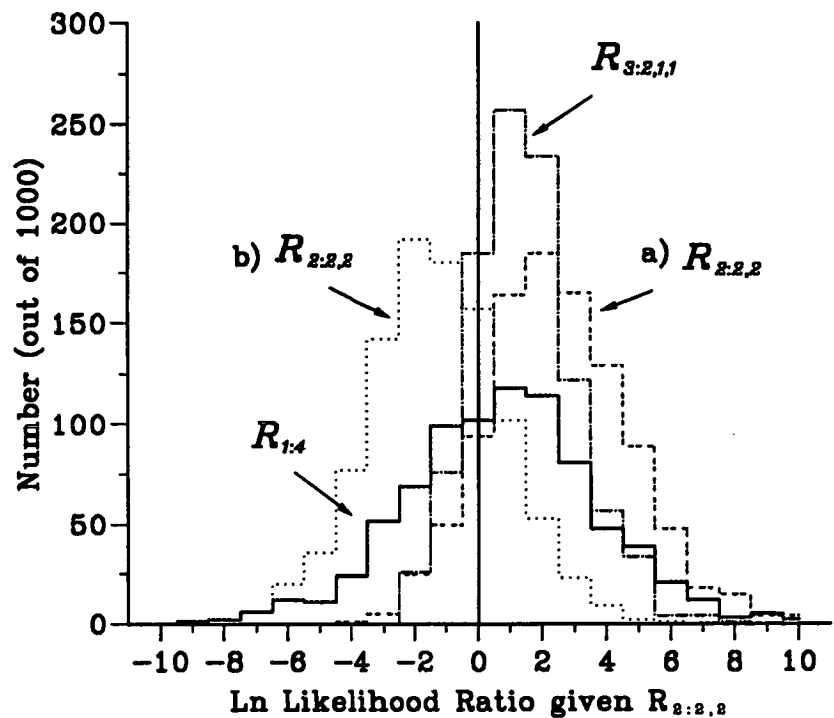


Figure 4.4b: The observed log likelihood ratios for each relationship, given that $R_{2:2,2}$ is the true relationship. The relationship marked a) has the true sibling pairs $R_{2:2,2}$ whereas b) includes the incorrect pairings.

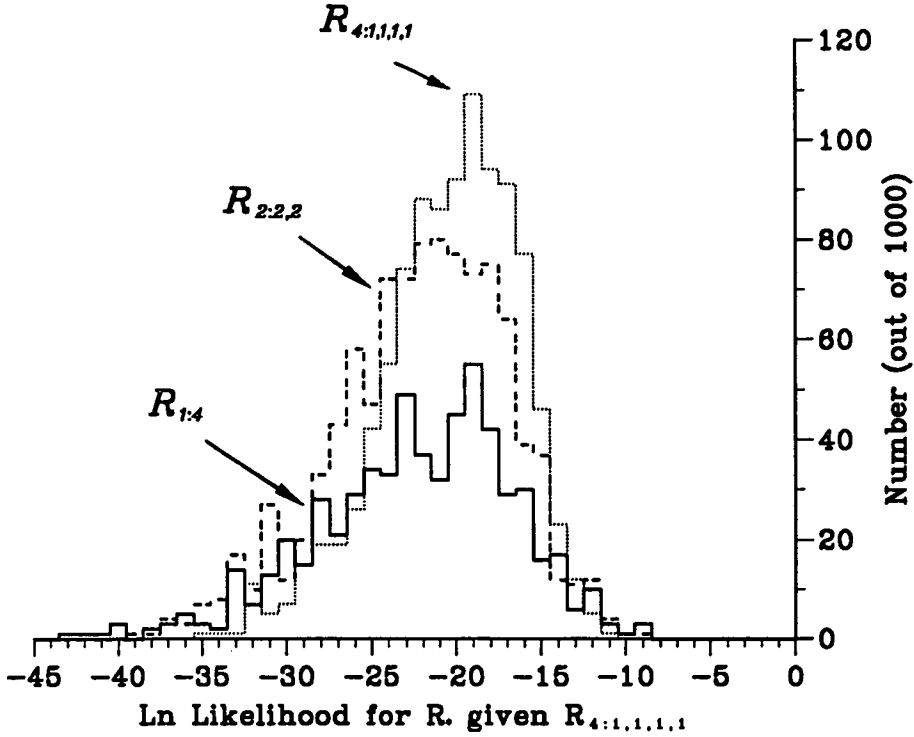


Figure 4.5a: The log likelihood for each relationship observed from simulation experiments for a sample of four individuals, given that $R_{4:1,1,1,1}$ is the true relationship.

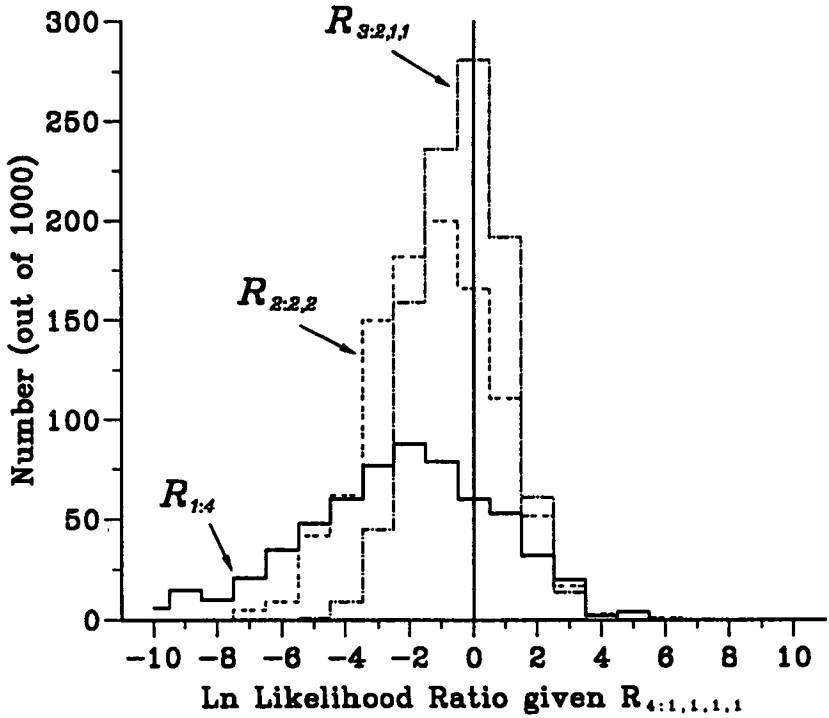


Figure 4.5b: The log likelihood ratios for each relationship, given four unrelated individuals. (The histogram for relationship $R_{2:3,1}$ lies between those for $R_{1:4}$ and $R_{2:2,2}$, and the histogram for $R_{3:2,1,1}$ in Figure 4.5a lies between $R_{2:2,2}$ and $R_{4:1,1,1,1}$ but are omitted for clarity.)

So, from these simulation results, it is seen that when the individuals are closely related, the true relationship tends to have the higher likelihood ratio values and the observed ratios are generally greater than one. If the individuals are less closely related, the mean likelihood ratios tend to be less than or close to one.

4.3.3 Likelihood Ratios for Relationships with Several Families

The likelihood ratios are calculated using the relationship $R_{N:1,1,\dots}$ as the standard, and so for samples with more than one family the ratios can often be simplified. For example, the likelihood ratio for $R_{2:2,1}$ is

$$\frac{\Pr(\Gamma_1, \Gamma_2, \Gamma_3 \mid R_{2:2,1} \text{ as } \{X_1, X_2\}\{X_3\})}{\Pr(\Gamma_1, \Gamma_2, \Gamma_3 \mid R_{3:1,1,1} \text{ as } \{X_1\}\{X_2\}\{X_3\})}$$

which is the same as the likelihood ratio for $R_{1:2}$,

$$\frac{\Pr(\Gamma_1, \Gamma_2 \mid R_{1:2} \text{ as } \{X_1, X_2\})}{\Pr(\Gamma_1, \Gamma_2 \mid R_{2:1,1} \text{ as } \{X_1\}\{X_2\})}$$

because the term for the family $\{X_3\}$ cancels out. Similarly, in Section 4.3.1, the expected likelihood ratios were found for samples of two and three individuals. The expected likelihood ratio for $R_{2:2,1}$ is the same as the expected likelihood ratio for $R_{1:2}$, given that at least two of the individuals are sibs.

Consider a sample of four individuals. The 15 relationship structures have been listed previously, but now consider the likelihood ratios.

- $R_{1:4}$: The relationship $R_{1:4}$ is unique and the likelihood ratio must be found as described previously.
- $R_{2:3,1}$: The likelihood ratios for the set of relationships $R_{2:3,1}$ are equivalent to the likelihood ratios for $R_{1:3}$ for each combination of three chosen from four, i.e. $\{X_1, X_2, X_3\}$, $\{X_1, X_2, X_4\}$, $\{X_1, X_3, X_4\}$ and $\{X_2, X_3, X_4\}$.
- $R_{3:2,1,1}$: Similarly, the set of relationships $R_{3:2,1,1}$ have likelihood ratios equal to those for the pairs $R_{1:2}$ for all possible pairs $\{X_1, X_2\}$, $\{X_1, X_3\}$, $\{X_1, X_4\}$, $\{X_2, X_3\}$, $\{X_2, X_4\}$, $\{X_3, X_4\}$.
- $R_{2:2,2}$: The likelihood ratios for the three relationships $R_{2:2,2}$ can be considered as the product of two pairs, and therefore

$$\frac{\Pr(\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4 \mid R_{2:2,2} \text{ as } \{X_1, X_2\}\{X_3, X_4\})}{\Pr(\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4 \mid R_{4:1,1,1,1} \text{ as } \{X_1\}\{X_2\}\{X_3\}\{X_4\})}$$

gives the same result as

$$\frac{\Pr(\Gamma_1, \Gamma_2 \mid R_{1:2} \text{ as } \{X_1, X_2\})}{\Pr(\Gamma_1, \Gamma_2 \mid R_{2:1,1} \text{ as } \{X_1\}\{X_2\})} \times \frac{\Pr(\Gamma_1, \Gamma_2 \mid R_{1:2} \text{ as } \{X_3, X_4\})}{\Pr(\Gamma_1, \Gamma_2 \mid R_{2:1,1} \text{ as } \{X_3\}\{X_4\})}$$

and similarly for the other two relationships $R_{2:2,2}$.

- $R_{4:1,1,1,1}$: The likelihood ratio for $R_{4:1,1,1,1}$ is always equal to one.

The pattern of likelihood ratios can therefore be built up by considering smaller subsets: all possible pairs, all possible triplets, etc. This could be incorporated into the analysis programs and would result in considerable savings in computing effort if the programs were to be used extensively for samples greater than six. For example, the likelihood ratio for the relationship $R_{6:2,1,1,1,1,1}$ is simply found by computing the likelihood ratio for relationship $R_{1:2}$ with the appropriate individuals as siblings.

4.4 Patterns of Inference

For any given sample of individuals, the likelihood for each relationship structure can be calculated; the relationship hypotheses are ranked in descending order based on the magnitude of the likelihood ratio for each relationship. In Chapter 3, the relationship structure ranked in first place is the one that is taken as the best estimate of the relatedness within that sample. In the first part of this chapter, it has been shown that the true relationship does not necessarily have the maximum expected likelihood ratio and so will not be consistently ranked first. So the question remains whether additional information regarding relatedness can be found by examining the relative order of the relationship hypotheses when ranked according to their likelihood ratios.

In order to detect trends in the pattern of ranked relationships, samples of four individuals were generated by the simulation program and analysed as described previously. The number of times each relationship is placed in *each* rank is recorded, rather than just the number of times each was ranked first. The results therefore form a square array; for $N = 4$, there are 15 relationship structures and therefore a maximum of 15 ranks into which each relationship may be placed. This array can be displayed as a three dimensional surface, or grid, to display trends in the pattern of ranked relationships.

Consider the case for $N = 4$ when the four are known to be full sibs $R_{1,4}$, or unrelated $R_{4:1,1,1,1}$. The set of grids in Figure 4.6 shows the pattern of inference with increasing quantities of data. In this example, each locus has two alleles with $p = (0.5, 0.5)$ so no relationships are excluded. The height of each bar in the grid represents the number of times a relationship is placed in a particular rank, taken over many replicate trials. For example, in Figure 4.6(a), the relationship $R_{1,4}$ is placed in the highest rank approximately 340 times. This is the simulated relationship and corresponds to an estimated probability of correct inference equal to 0.34 if inference is based on first rank only. The row of bars in each grid for the first rank inference shows the results described in Chapter 3.

When data are available from only one locus (Figure 4.6(a) and 4.6(c)), the number of relationships placed in each rank is affected by “draws”, visible here as high bars in only a few positions. Draws tend to occur among members of the same general group, for example the three structures within $R_{2,2,2}$, so that in the case of four individuals draws will tend to occur at ranks two, six and nine. As the number and variety of loci increases, however, the number of draws decreases (see Section 4.2.2). With information available from many loci (Figure 4.6(b) and 4.6(d)), a distinct pattern can be seen. When the individuals are closely related, relationship structures that include larger sibships tend to be placed in the higher ranks. This is seen as a ridge running from left to right of the grid in Figure 4.6(b). Conversely, if the individuals are not closely related then the reverse is seen and relationship structures with large sibships appear more often in the lower ranks; the ridge runs from top to bottom of the grid in Figure 4.6(d). When relationship $R_{4:1,1,1,1}$ is the simulated relationship, it is most often placed in the second rank position when there are data from ten loci.

Note that in this example, when the true relationship is either $R_{1,4}$ or $R_{4:1,1,1,1}$, relationship structures within the main groups are equivalent in the long run. For example the same pattern of ranking is seen for each of the relationships $R_{2,3,1}$ when taken over the 1000 replicate samples.

The example shown in Figure 4.6 is not affected by exclusion as each locus has two alleles. For an example in which exclusion was possible, there would be a different pattern; the peaks in the graph would be shifted towards the higher ranks as fewer valid relationships would be counted in each trial.

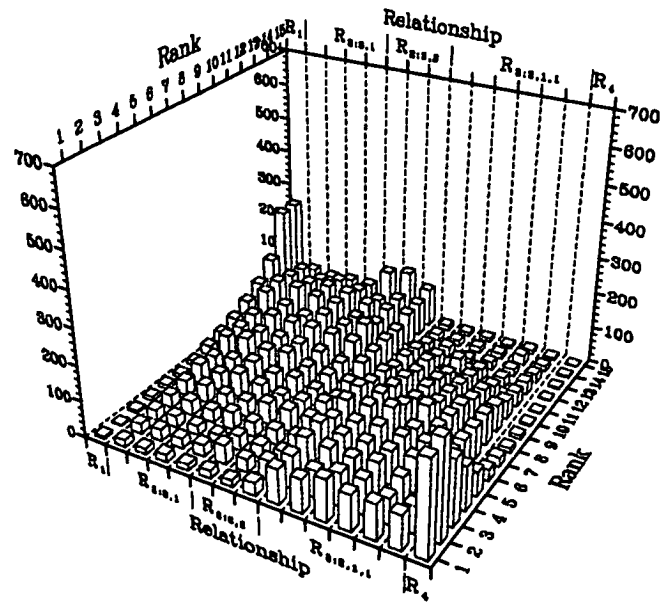
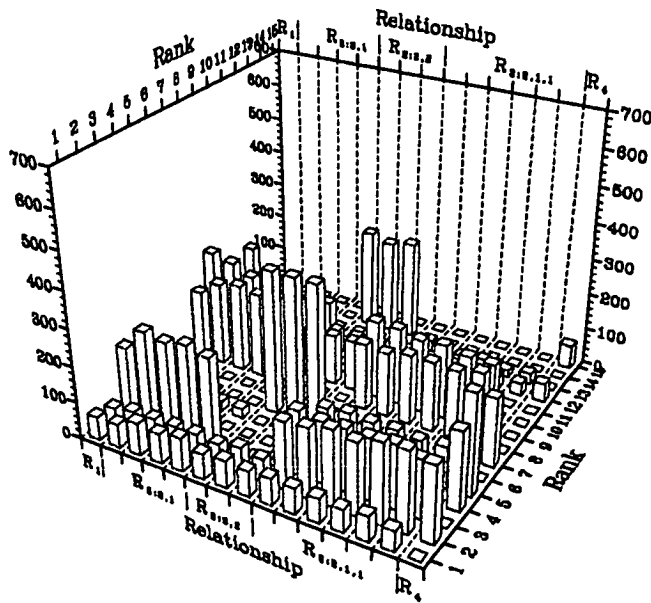
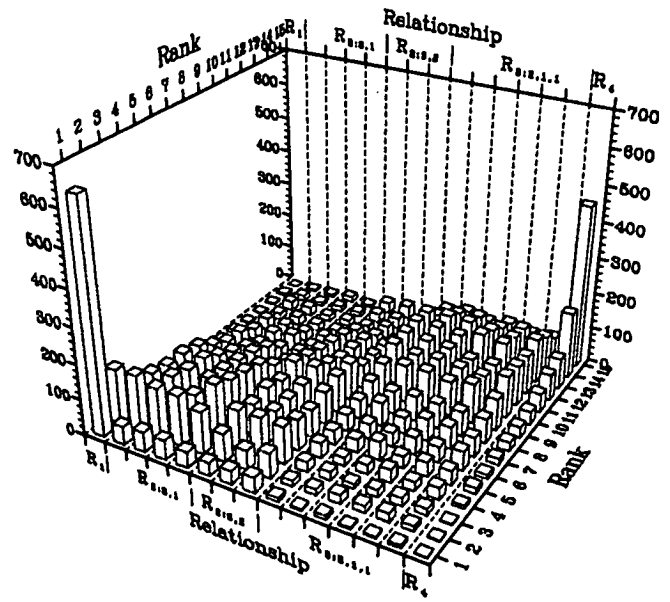
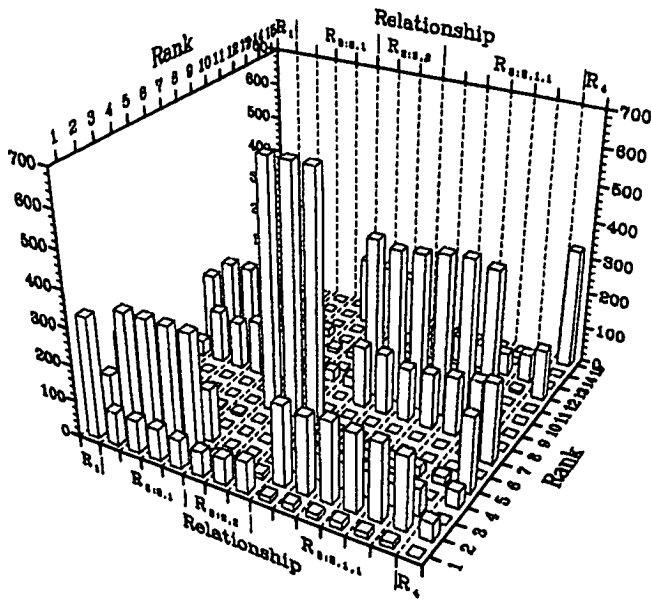


Figure 4.6: The number of relationships placed in each rank for a sample of four individuals with genotype data from one or ten loci. Each locus has two alleles with equal frequency and there are 1000 replicates.

If no information regarding relationship could be obtained by this method of analysis, then an approximately flat surface would be expected for the grids described above; each relationship could be ranked in any position with roughly equal frequency. This is not seen, particularly when the individuals in the sample are closely related to each other, although it is possible for any relationship structure to be ranked in any position when taken over many trials. It may be possible to calculate some value to indicate the “power” of the method by comparing the observed results from simulation with the “null hypothesis” of a flat surface. For the examples above, with no exclusion each relationship would be expected to occur in each rank on two out of three occasions if there was no information in the likelihood ratios. Figure 4.6 shows how typical results differ from this. This approach would be affected by exclusion, however, and a different way to record draws would be required.

Field Parameters

Taking a more realistic set of locus parameters, Figure 4.7 shows the number of times that each relationship is ranked in each position using the parameters for the seven field loci (see Section 3.2.4). In this example there are 10,000 replicate trials. Similar patterns can be found for each of the major relationship structures. The results obtained from simulation experiments are shown in Figures 4.8 to 4.11 for $R_{2:3,1}$, $R_{2:2,2}$, $R_{3:2,1,1}$ and $R_{4:1,1,1,1}$ respectively. Note that the same scale is used for all plots so that the patterns are directly comparable.

In each case except $R_{4:1,1,1,1}$, the true relationship is the one ranked in first position most often; if the correct relationship is not first then it tends to be ranked second or third. Note that every relationship is placed in every rank at least once; but some patterns of inference do emerge.

1. $R_{1:4}$, Figure 4.7: When the individuals are all full sibs, the four different relationships $R_{2:3,1}$ are predominantly placed in the ranks 2–5; each of the three relationships $R_{2:2,2}$ can appear in any rank but tend to be placed in the middle ranks; the six relationships $R_{3:2,1,1}$ are placed in the lowest ranks with $R_{4:1,1,1,1}$ generally placed last.

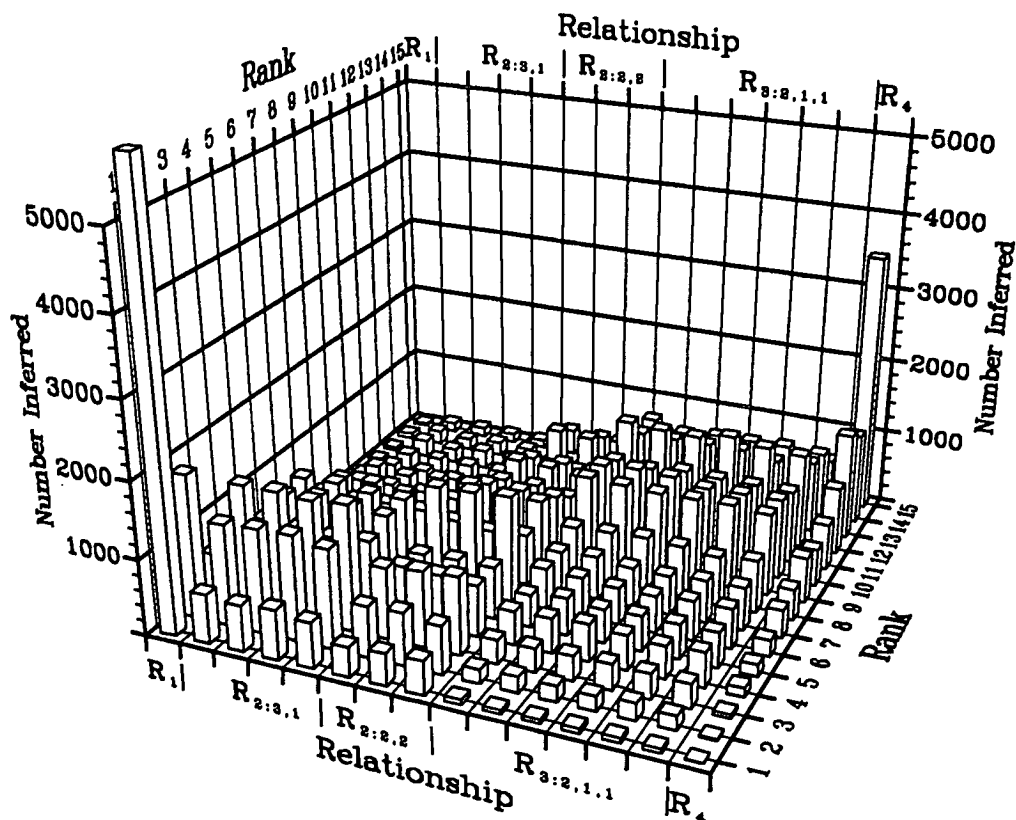


Figure 4.7: The Pattern of Inference in all ranks for a sample of four full siblings, using field parameters. There are 10,000 replicates.

2. $R_{2:3,1}$, Figure 4.8: When this is the simulated relationship (as $\{X_1X_2X_3\}\{X_4\}$ in this example) then the correct relationship structure will tend to be ranked first; $R_{1,4}$ is also generally placed in the higher ranks; the alternative relationships $R_{2:3,1}$ are predominantly placed in the lowest ranks and the relationships $R_{2:2,2}$ and $R_{4:1,1,1,1}$ are placed anywhere in the final 10 ranks. The six relationships $R_{3:2,1,1}$ are divided into two groups. Those in which the pair are true sibs (any two from X_1, X_2 or X_3) are placed in the top few ranks while those that do not are placed in the lowest ranks.
3. $R_{2:2,2}$, Figure 4.9: When $R_{2:2,2}$ is the true relationship (as $\{X_1X_2\}\{X_3X_4\}$ here) the relationships $R_{1,4}$ and $R_{2:3,1}$ seem to be placed equally frequently in any rank; the correct relationship structure $R_{2:2,2}$ is placed in the highest ranks, as are the two relationships $R_{3:2,1,1}$ that have a correct pair of sibs as a subset; the incorrect

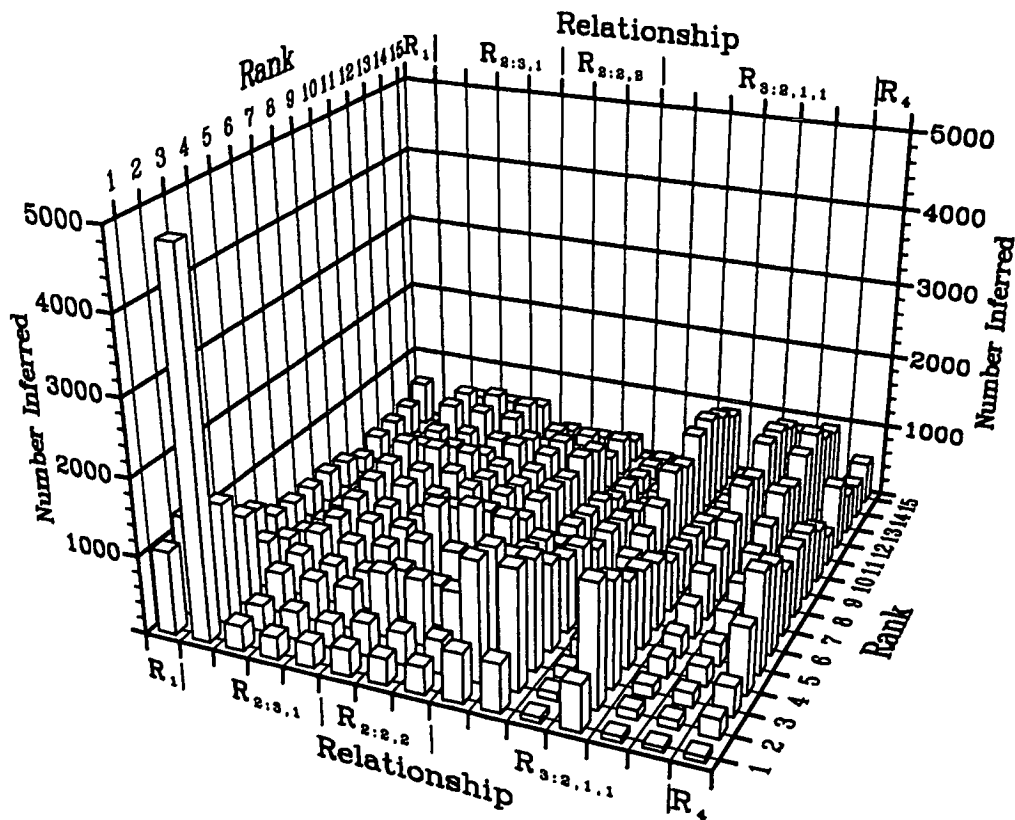


Figure 4.8: The Pattern of Inference in all ranks for a sample of four individuals, using field parameters, given $R_{2:3,1}$. There are 10,000 replicates.

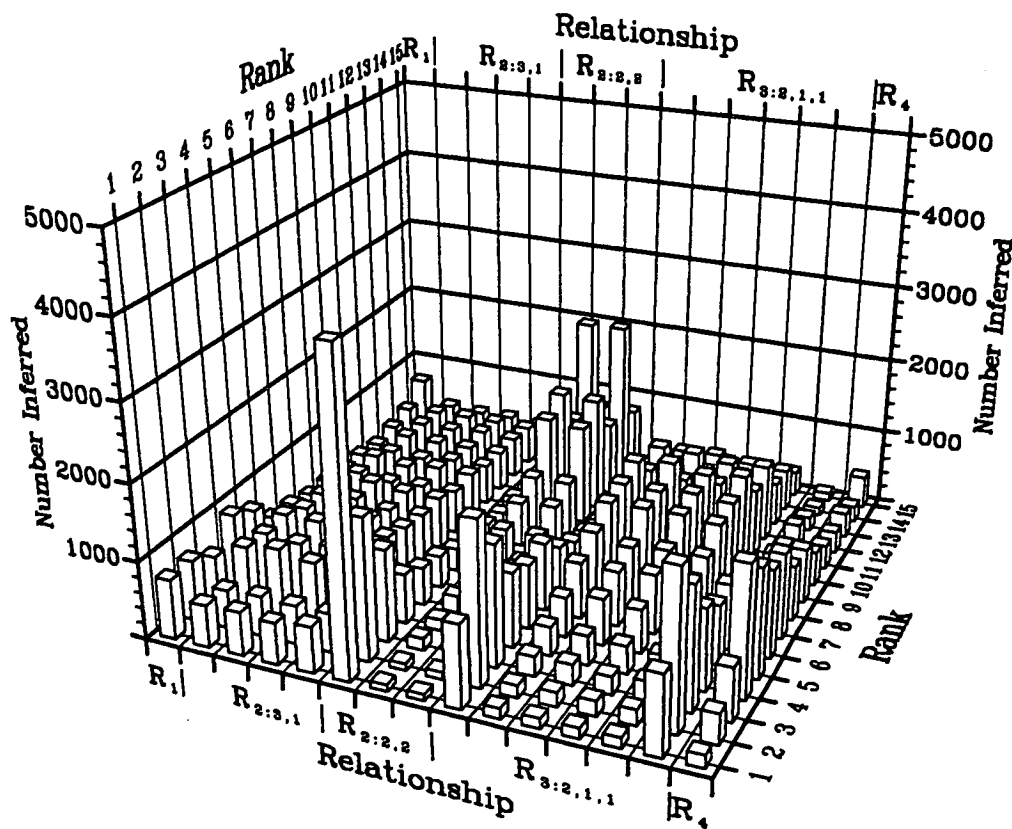


Figure 4.9: The Pattern of Inference in all ranks for a sample of four individuals, using field parameters, given $R_{2:2,2}$. There are 10,000 replicates.

$R_{2:2,2}$ relationships are placed last behind the remaining four $R_{3:2,1,1}$ structures.

$R_{4:1,1,1,1}$ tends to appear in the middle ranks.

4. $R_{3:2,1,1}$, Figure 4.10: When the simulated relationship is $R_{3:2,1,1}$ (as $\{X_1X_2\}\{X_3\}\{X_4\}$ here), this relationship structure is often placed in the highest rank; the relationship $R_{1:4}$ and the relationships $R_{2:3,1}$ and $R_{2:2,2}$ that do not have the correct pair as sibs tend to appear most often in the lower ranks; the other relationships that *do* have the correct pair as sibs can appear in any rank but with a slight bias towards the higher ranks. The relationship structures $R_{3:2,1,1}$ with the incorrect pairing are generally placed in the middle ranks, while $R_{4:1,1,1,1}$ is placed predominantly in one of the higher ranks.
5. $R_{4:1,1,1,1}$, Figure 4.11: If the four individuals are unrelated then any relationship can appear in any rank with about equal frequency. The correct relationship does, however, occur most often in the second rank and usually in the first seven ranks; $R_{1:4}$ tends to come last. There is a pattern visible such that the more closely related the individuals the lower that relationship is ranked, but the pattern is less well defined than in the previous examples.

Larger Samples

For a sample of four individuals, patterns or trends are seen in the position of relationships in the ranked list of likelihood ratios. Relationship structures that are similar to the true relationship tend to be placed high in the ranked list, whereas those relationship structures that are very different from the true relationship are placed near the bottom of the ranked list. These general patterns are also seen for larger sample sizes. When the individuals are closely related the correct relationship is the one most often placed first, followed by those relationship structures that have a correct sibship as a subset (either a whole family or a part of a true family). Relationships that split up a true family tend to be placed in the lower ranks. When the individuals are less closely related, the pattern of relationships in ranks is less well defined but taken over many trials, relationship structures that consider the individuals as sibs are ranked lower than those that do not.

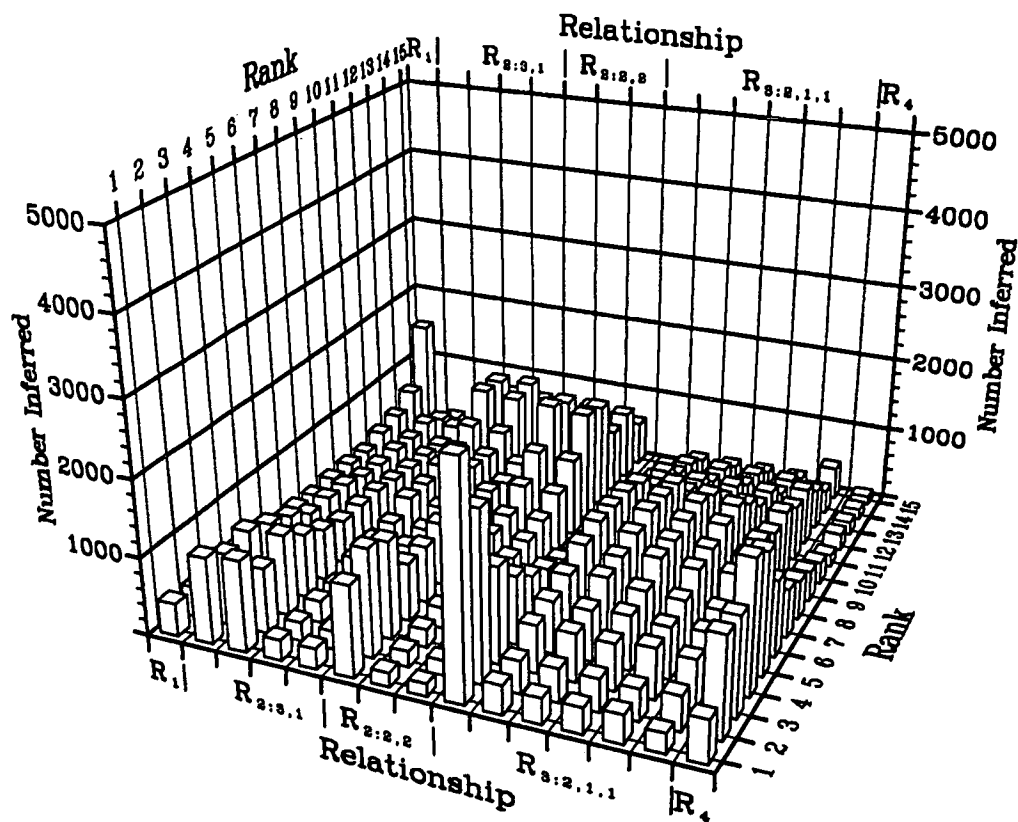


Figure 4.10: The Pattern of Inference in all ranks for a sample of four individuals, using field parameters, given $R_{3:2,1,1}$. There are 10,000 replicates.

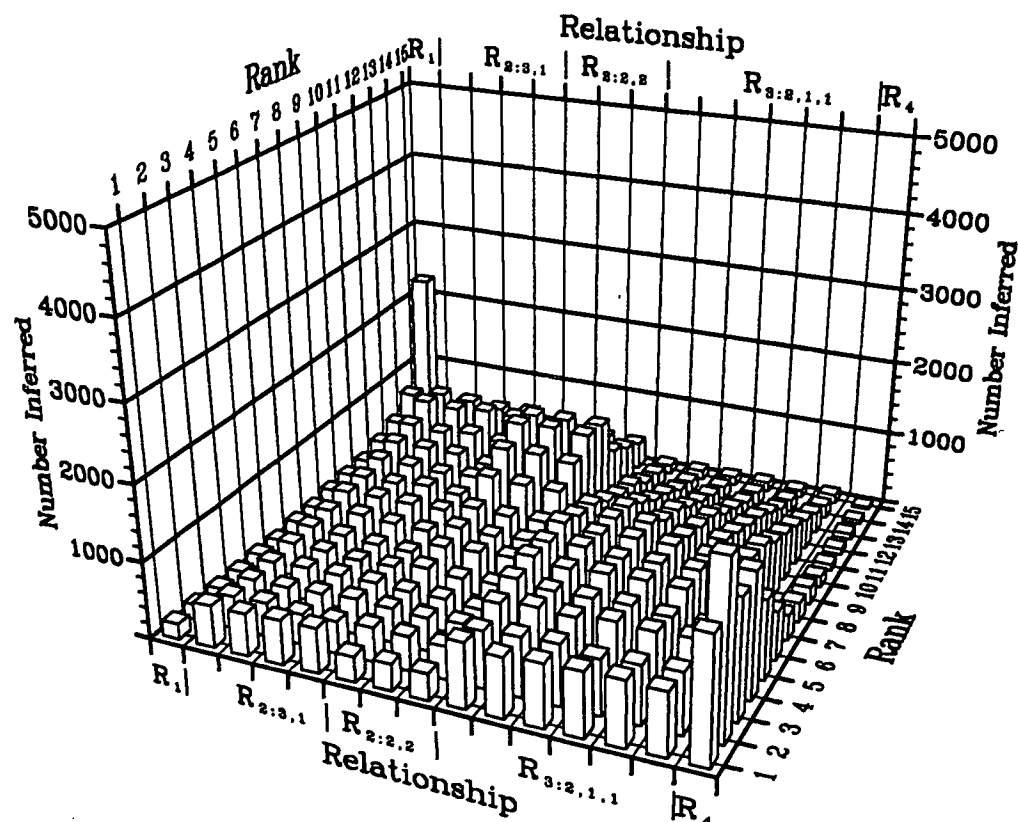


Figure 4.11: The Pattern of Inference in all ranks for a sample of four individuals, using field parameters, given $R_{4:1,1,1,1}$. There are 10,000 replicates.

As an example, the two extreme cases for a sample of five individuals are given in Figures 4.12a ($R_{1:5}$) and 4.12b ($R_{5:1,1,1,1,1}$) and clearly show the pattern described above. Only one example from each main group of relationships is shown for clarity, instead of all 52 possible relationships. Note that when $R_{5:1,1,1,1,1}$ is the true relationship, this is placed in the first rank only 3% of the time. For six individuals and 10,000 replicate samples the same general pattern is seen.

4.5 Summary

In this chapter, the likelihood ratios for each relationship are sorted into descending order of magnitude and the pattern of relationships within this list is examined.

Relationships for which the likelihood is zero are said to be excluded. It has been shown that the excluding power of a locus, or set of loci, can be found by estimating the probability that a randomly chosen set of individuals have genotypes that are compatible for a particular relationship. The excluding power of a locus is zero if there are only two alleles or if the relationship structure has only families of two siblings. For a given sample size, the excluding power of a locus is maximum when the alleles have equal frequency. Exclusion alone cannot determine the true relationship between individuals: there will always be more than one relationship structure for which the likelihood is greater than zero. For single samples, however, some useful information may be found by examining the type of excluded relationships. For example, the true relationship and those that contain true subsets of families cannot be excluded.

For relationships that are not excluded by the set of genotypes observed in the sample, the expected likelihood ratios can be found or estimated from simulation experiments. It has been shown that the true relationship does not necessarily give the maximum expected likelihood ratio, and hence the method appears not to be consistent. When the individuals are very closely related the true relationship does have the maximum expected likelihood ratio, but when the individuals are less closely related this is no longer the case. The expected likelihood ratio is maximal for the true hypothesis only when comparisons are made between relationships for a given set of individuals. However, within the set of relationship structures considered for N individuals, there

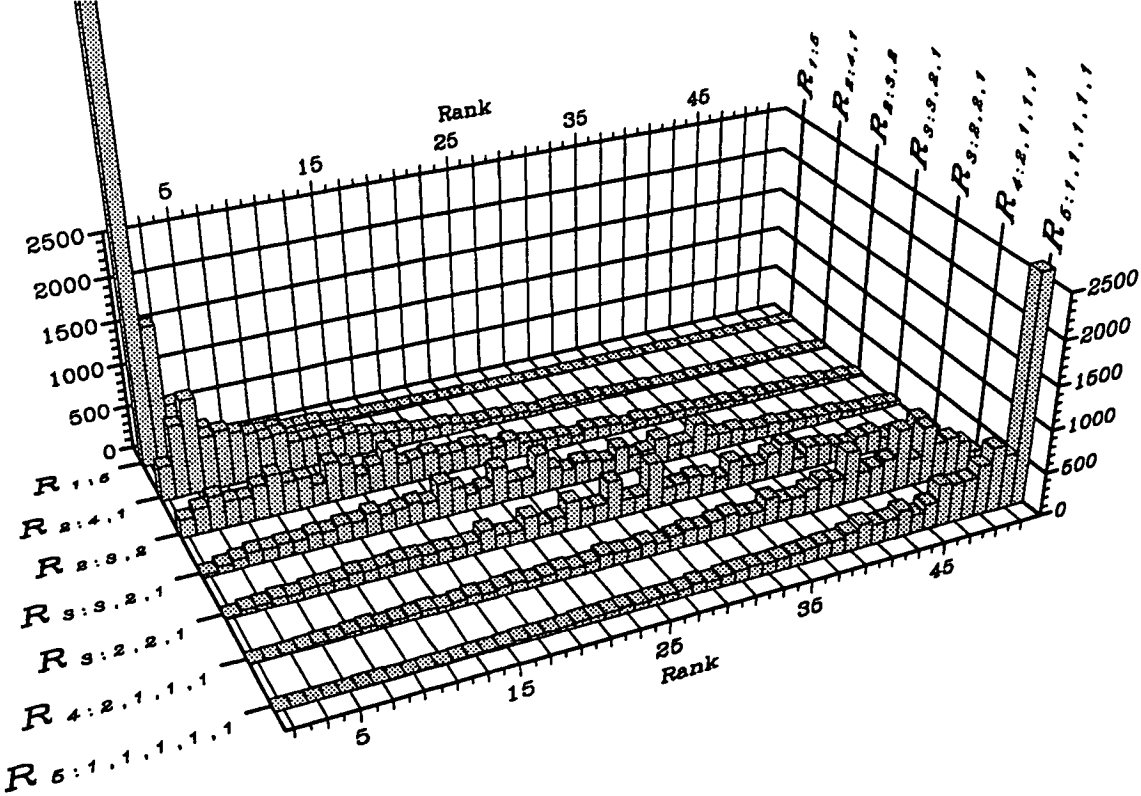


Figure 4.12a: The number of samples placed in each rank for each relationship, given a family of five individuals. Note that the number of samples in which $R_{1:5}$ was ranked first is 5326 but the graph is truncated in order to show the same scale as Figure 4.12b.

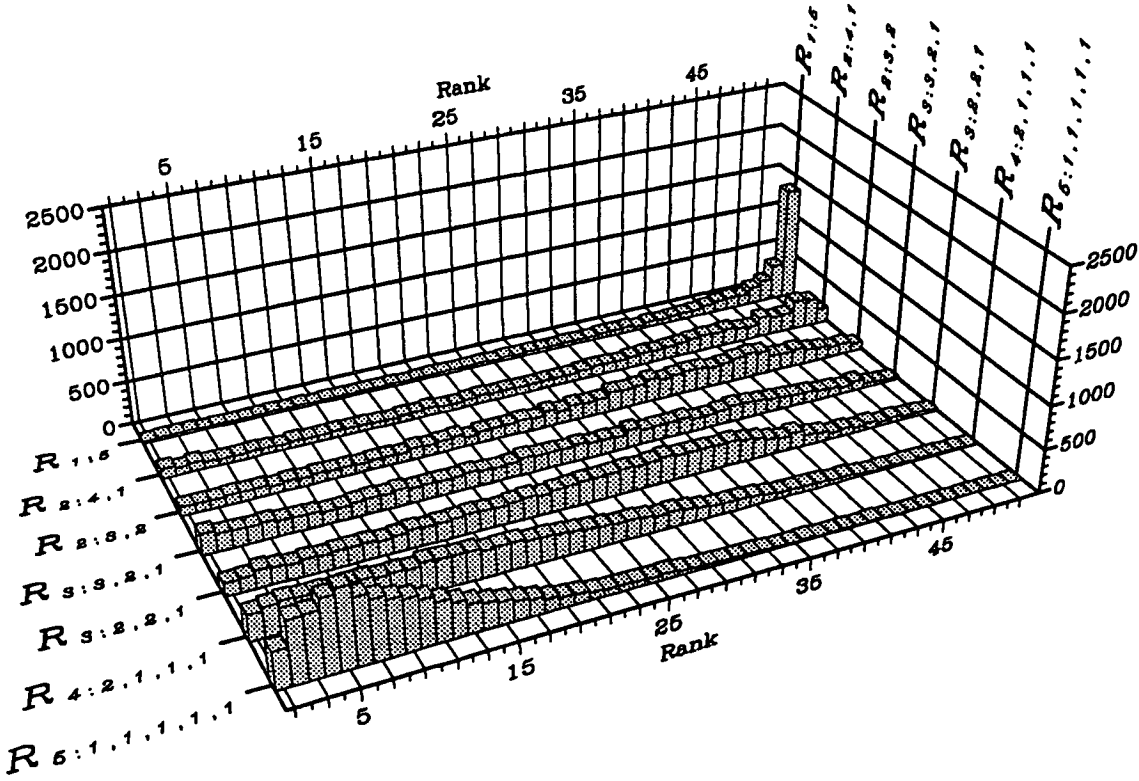


Figure 4.12b: The number of samples placed in each rank for each relationship, given $R_{5:1,1,1,1,1}$.

are different types of hypothesis; alternative hypotheses may compare individuals for given relationships rather than relationships between a given set of individuals. This is discussed in Chapter 7.

Using only the highest ranked hypothesis to estimate the genetic relationship between individuals has been shown to be unreliable. Useful information is lost by ignoring trends in the ordered list of the likelihood ratios for each relationship structure. It has been shown in this chapter that the true relationship and relationships that are similar to it will be placed in the higher ranks, while structures that split up true sibships will be placed in lower ranks. If the individuals are closely related, this is revealed by the fact that relationship structures that contain true (and therefore similar) families will be grouped together in the ranked list of likelihood ratios. When there are data from several loci, if the individuals are not closely related then the order of relationships in the ranked list does not show this grouping of similar relationships.

In Chapter 5, some examples are given showing the interpretation of single sample data using the methods developed in Chapters 3 and 4.

Chapter 5

The Method Applied to Single Samples

5.1 Introduction

In Chapters 3 and 4 different methods of interpreting the ranked likelihood ratios were presented, and applied to simulated data. Although it is useful to evaluate the method by applying it to samples with known structure, using simulated data, it is essential to consider the results obtained when the method is applied to *single* samples in which the true structure is unknown. The interpretation of the likelihood for each relationship found for a single sample is discussed in this chapter.

In order to draw some useful conclusions regarding the relatedness within a single sample, there are three main features of interest.

1. The total number of possible relationships, given certain assumptions regarding mating patterns or known species behaviour.
2. The number of these relationships that are valid; i.e. the number of the possible relationships that are not excluded on the basis of the observed genotypes in the sample (or excluded on any other grounds such as maximum family size).
3. The range and variability of the likelihoods for the subset of valid relationship structures.

Each of these features is now considered for eight examples. The first set of examples given in Section 5.2 show the analysis carried out for four samples where each has

been simulated with a known relationship structure and known population parameters. These samples were provided by Professor W.G. Hill and analysed by me without prior knowledge of the simulated relationship. After conclusions were reached regarding the relationship within each sample, the true relationship was disclosed and the answers compared. The second set of examples, in Section 5.3, were taken from Quinn *et al* [1987]. These data are analysed using the likelihood method and the conclusions compared with those of the original authors.

5.2 Analysis of Single Simulated Samples

Example A

Sample Genotypes: These are the genotypes and allele frequencies for this sample of four individuals. There are eight loci, each with two alleles.

Locus number and p								
	1	2	3	4	5	6	7	8
X	(0.2,0.8)	(0.6,0.4)	(0.4,0.6)	(0.8,0.2)	(0.2,0.8)	(0.6,0.4)	(0.4,0.6)	(0.8,0.2)
1	A_2A_2	A_2A_1	A_1A_1	A_1A_1	A_2A_1	A_1A_2	A_2A_2	A_2A_1
2	A_2A_2	A_1A_2	A_2A_1	A_1A_1	A_1A_2	A_2A_2	A_2A_1	A_1A_1
3	A_2A_2	A_2A_2	A_2A_2	A_2A_1	A_2A_2	A_2A_1	A_2A_1	A_1A_1
4	A_2A_2	A_1A_1	A_1A_2	A_1A_1	A_1A_1	A_2A_2	A_2A_1	A_1A_1

As there are four individuals in this sample, there are 15 possible relationships to consider. As each locus has only two alleles segregating there cannot be any exclusions. The result of the analysis for all loci combined is given in Table 5.1. The log likelihood ratios for each relationship are listed in descending order of magnitude. With data available from eight independent loci, the probability of observing these exact genotypes for a set of four individuals is very small. The likelihood for the relationship $R_{3:2,1,1}$ as $\{X_2, X_4\}\{X_1\}\{X_3\}$ is 1.3258×10^{-12} ; the log likelihood ratio for this relationship relative to $R_{4:1,1,1,1}$ is 2.3378.

Solution 1: If the likelihoods for each relationship are combined over all loci, the relationship structure that gives the highest probability for the data is $R_{3:2,1,1}$ as

Rank	Relationship		Log Likelihood Ratio
	R_*	as	
1	$R_{3:2,1,1}$	$\{X_2, X_4\}\{X_1\}\{X_3\}$	2.3378
2	$R_{2:3,1}$	$\{X_1, X_2, X_4\}\{X_3\}$	1.0502
3	$R_{3:2,1,1}$	$\{X_1, X_2\}\{X_3\}\{X_4\}$	0.0739
4	$R_{4:1,1,1,1}$	$\{X_1\}\{X_2\}\{X_3\}\{X_4\}$	0.0000
5	$R_{3:2,1,1}$	$\{X_1, X_4\}\{X_2\}\{X_3\}$	-0.7768
6	$R_{2:2,2}$	$\{X_1, X_3\}\{X_2, X_4\}$	-0.8221
7	$R_{2:3,1}$	$\{X_2, X_3, X_4\}\{X_1\}$	-1.0057
8	$R_{3:2,1,1}$	$\{X_2, X_3\}\{X_1\}\{X_4\}$	-1.0962
9	$R_{2:2,2}$	$\{X_1, X_4\}\{X_2, X_3\}$	-1.8730
10	$R_{2:2,2}$	$\{X_1, X_2\}\{X_3, X_4\}$	-3.0860
11	$R_{3:2,1,1}$	$\{X_1, X_3\}\{X_2\}\{X_4\}$	-3.1599
11	$R_{3:2,1,1}$	$\{X_3, X_4\}\{X_1\}\{X_2\}$	-3.1599
13	$R_{2:3,1}$	$\{X_1, X_2, X_3\}\{X_4\}$	-3.2158
14	$R_{1:4}$	$\{X_1, X_2, X_3, X_4\}$	-3.9794
15	$R_{2:3,1}$	$\{X_1, X_3, X_4\}\{X_2\}$	-5.2795

Table 5.1: Results for Example A, showing the log likelihood ratio for each of the 15 possible relationships for the four individuals X_1 – X_4 . These are the combined totals for eight loci; each locus has two alleles.

$\{X_2, X_4\}\{X_1\}\{X_3\}$, i.e. three families with individuals X_2 and X_4 as siblings (see Table 5.1). This is the relationship hypothesis that gives the best estimate of the true relationship between these four individuals, based on the first rank only.

Solution 2: The expected pattern of inference for a set of seven loci and four individuals is described in Section 4.4. The same procedure was repeated here; the parameters for the eight loci given above were used to randomly generate genotypes for samples of four individuals with known relatedness (1000 for each of the relationships $R_{1:4}$, $R_{2:3,1}$, $R_{2:2,2}$, $R_{3:2,1,1}$ and $R_{4:1,1,1,1}$). These simulated samples were then analysed and the inference pattern recorded. The profile of results from the analysis of these samples was found to be very similar to the previous patterns found in Section 4.4. The following table shows the number of times each relationship was placed in each rank. Note that not all 15 possible relationships are shown for every case as some relationships are equivalent; for example, if the simulated relationship is $\{X_1, X_2, X_3\}\{X_4\}$ (represented as $\{iii\}\{j\}$), then the relationships $\{X_1, X_2, X_4\}\{X_3\}$, $\{X_1, X_3, X_4\}\{X_2\}$

and $\{X_2, X_3, X_4\}\{X_1\}$ all give very similar results in the long run and are therefore represented as $\{ijj\}\{i\}$ in the table below. In each case, the simulated relationship is placed in rank one more often than any other relationship but the proportion of correct inferences decreases with decreasing relatedness, as found previously.

Relationship	No. of times Relationship placed in rank -														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
{iii} simulated															
{iii}	578	167	56	54	33	23	15	17	11	13	9	11	2	5	6
{iii}{i}	62	125	134	114	113	102	76	68	45	52	31	25	26	18	10
{ii}{ii}	39	63	102	93	92	95	96	106	67	53	51	44	31	38	32
{ii}{i}{i}	9	24	26	38	41	48	57	63	89	111	101	122	114	89	70
{i}{i}{i}{i}	5	5	8	13	17	25	29	32	35	40	52	65	94	175	405
{iii}{j} simulated															
{iii}{j}	85	115	51	75	81	65	57	60	72	48	57	52	56	63	63
{iii}{j}	523	158	86	65	36	27	32	23	14	16	4	9	4	1	2
{ij}{i}	22	21	44	42	46	46	66	77	75	97	93	97	105	96	72
{ii}{ij}	37	41	73	61	91	91	93	91	85	75	75	60	47	40	39
{ii}{i}{j}	62	167	149	132	94	79	67	57	40	47	30	25	21	19	11
{ij}{i}{i}	6	14	19	27	30	38	50	62	71	83	90	109	124	133	150
{i}{i}{i}{j}	12	25	39	72	132	142	80	71	82	66	40	64	48	74	53
{ii}{j}{j} simulated															
{ii}{j}{j}	65	52	53	86	77	49	64	61	60	84	71	62	66	72	78
{ii}{j}{j}	39	36	76	62	78	75	83	81	76	77	79	73	66	58	41
{ii}{j}{j}	472	190	96	61	39	31	20	20	19	24	8	5	8	4	3
{ij}{ij}	8	13	12	15	21	30	24	44	55	76	74	111	122	219	180
{ii}{j}{j}	108	242	181	98	73	68	49	40	36	25	19	24	18	9	13
{ij}{i}{j}	15	17	23	40	70	87	84	97	96	90	89	98	74	71	52
{i}{i}{j}{j}	16	42	81	230	116	100	86	72	48	43	32	27	35	37	35
{ii}{j}{k} simulated															
{ii}{j}{k}	25	20	30	38	44	27	48	51	64	60	65	97	96	139	196
{ij}{k}	92	94	94	75	71	66	69	68	72	70	56	60	51	38	30
{ij}{k}{i}	10	13	21	22	26	39	36	64	70	99	84	108	139	156	115
{ii}{j}{k}	146	149	135	83	64	75	50	61	52	42	39	42	26	20	16
{ij}{ik}	14	21	21	34	40	50	43	63	67	89	110	90	114	113	131
{ii}{j}{k}	361	228	120	68	64	38	32	22	21	15	6	12	4	4	5
{ij}{i}{k}	36	41	63	75	84	97	101	89	81	74	70	63	50	43	34
{i}{i}{j}{k}	60	149	140	177	154	84	52	41	32	28	19	20	16	17	11
{i}{j}{k}{l} simulated															
{ijk}{l}	8	4	8	17	20	9	30	27	31	55	58	89	115	191	338
{ijk}{l}	27	36	38	36	40	51	65	71	71	85	92	100	104	109	73
{ij}{kl}	42	42	37	44	50	59	58	60	75	87	96	84	92	85	92
{ij}{k}{l}	94	82	93	89	95	92	84	83	72	59	49	42	30	25	12
{i}{j}{k}{l}	184	224	171	173	93	52	31	18	20	13	6	8	4	2	1

Returning to the single sample of four individuals in Example A, note that $R_{1,4}$ is placed last and $R_{4,1,1,1}$ is placed fourth. One of the relationships $R_{3,2,1,1}$ is placed first while the other $R_{3,2,1,1}$ are spread out in the other ranks. One relationship $R_{2,3,1}$ is placed second but the others are ranked much lower. The relationships $R_{2,2,2}$ are found in the middle ranks. This pattern fits well with that expected when the true relationship is $R_{3,2,1,1}$ as $\{X_2, X_4\}\{X_1\}\{X_3\}$. This confirms the previous conclusion that $R_{3,2,1,1}$ as $\{X_2, X_4\}\{X_1\}\{X_3\}$ is the best estimate of the true relationship between the individuals

in this sample, i.e. three families with individuals X_2 and X_4 full sibs in one family and the other two individuals unrelated.

Correct Solution The true relationship between the four individuals in Example A is $R_{3:2,1,1}$ as $\{X_2, X_4\}\{X_1\}\{X_3\}$. So, for this example, the relationship between the four individuals has been inferred correctly.

Example B

Sample Genotypes: For sample B, there are also four individuals, but with genotypes known for five loci. Each locus has three alleles with frequencies 0.2, 0.3 and 0.5. Although exclusions would be possible with these parameters the likelihood for each relationship is greater than zero with the particular genotypes observed in sample B.

Locus number and $\underline{p}$					
	1	2	3	4	5
X	(0.2,0.3,0.5)	(0.2,0.3,0.5)	(0.2,0.3,0.5)	(0.2,0.3,0.5)	(0.2,0.3,0.5)
1	A_3A_2	A_1A_1	A_1A_3	A_3A_2	A_3A_3
2	A_2A_3	A_2A_3	A_1A_3	A_1A_2	A_3A_2
3	A_3A_3	A_3A_1	A_2A_3	A_3A_3	A_1A_3
4	A_2A_3	A_1A_3	A_2A_3	A_1A_2	A_3A_2

Solution 1: Combining the likelihoods for the five loci, the highest ranked relationship in this sample is found to be $R_{3:2,1,1}$ as $\{X_2, X_4\}\{X_1\}\{X_3\}$ (see Table 5.2). This is therefore the best estimate of the relatedness in the sample.

Solution 2: The simulation experiment is carried out as before, using the parameters appropriate for these five loci; the number of times each relationship is placed in each rank is recorded. The overall results are very similar to those shown in both Section 4.4 and Example A above and so are not given in detail here. In contrast to Example A, however, simulations show that with the parameters given for Example B, relationship $R_{1:4}$ would be excluded in approximately half of the trials. The relationships $R_{2:3,1}$ would be expected to be excluded in approximately 18% and 30% of samples (depending on whether the true sibs were included in the family of size three). Although there are

Rank	Relationship		Log Likelihood Ratio
	R_*	as	
1	$R_{3:2,1,1}$	$\{X_2, X_4\}\{X_1\}\{X_3\}$	0.9493
2	$R_{2:3,1}$	$\{X_1, X_2, X_4\}\{X_3\}$	0.8574
3	$R_{4:1,1,1,1}$	$\{X_1\}\{X_2\}\{X_3\}\{X_4\}$	0.0000
4	$R_{2:2,2}$	$\{X_1, X_3\}\{X_2, X_4\}$	-0.2014
5	$R_{3:2,1,1}$	$\{X_1, X_4\}\{X_2\}\{X_3\}$	-0.4212
6	$R_{3:2,1,1}$	$\{X_1, X_2\}\{X_3\}\{X_4\}$	-0.6548
7	$R_{3:2,1,1}$	$\{X_3, X_4\}\{X_1\}\{X_2\}$	-0.9425
8	$R_{3:2,1,1}$	$\{X_1, X_3\}\{X_2\}\{X_4\}$	-1.1507
9	$R_{2:3,1}$	$\{X_2, X_3, X_4\}\{X_1\}$	-1.5027
10	$R_{2:2,2}$	$\{X_1, X_2\}\{X_3, X_4\}$	-1.5973
11	$R_{2:3,1}$	$\{X_1, X_3, X_4\}\{X_2\}$	-2.7757
12	$R_{1:4}$	$\{X_1, X_2, X_3, X_4\}$	-3.1438
13	$R_{3:2,1,1}$	$\{X_2, X_3\}\{X_1\}\{X_4\}$	-3.7534
14	$R_{2:2,2}$	$\{X_1, X_4\}\{X_2, X_3\}$	-4.1746
15	$R_{2:3,1}$	$\{X_1, X_2, X_3\}\{X_4\}$	-4.7216

Table 5.2: The ranked log likelihood ratios for each relationship considered for the individuals in Example B. These are the combined totals for five loci; each locus has three alleles.

no exclusions in Example B, the relative order of the relationships in the ranking agrees with that expected when the true relationship is $R_{3:2,1,1}$, as $\{X_2, X_4\}\{X_1\}\{X_3\}$.

Correct Solution The true relationship between the four individuals in Example B is $R_{3:2,1,1}$ as $\{X_2, X_4\}\{X_1\}\{X_3\}$ and so the true relationship has been correctly inferred by both approaches.

Example C

Sample Genotypes: For Example C, there are only three individuals with genotypes known for seven loci; five of the loci have two alleles and the other loci have three alleles. As there are three individuals there are only five possible relationships to consider. The likelihood for each relationship is greater than in the previous examples as there are fewer individuals, but the log likelihood ratios are of similar magnitude (see Table 5.3).

The relationship $R_{1:3}$ is not excluded on the basis of the observed genotypes in this sample, so all five relationships are valid.

X	Locus number and p						
	1	2	3	4	5	6	7
	(0.2,0.8)	(0.6,0.4)	(0.4,0.6)	(0.4,0.6)	(0.8,0.2)	(0.2,0.3,0.5)	(0.2,0.3,0.5)
1	A_1A_2	A_1A_2	A_1A_2	A_2A_2	A_1A_2	A_2A_3	A_2A_3
2	A_2A_2	A_1A_2	A_1A_2	A_1A_1	A_1A_1	A_2A_3	A_1A_3
3	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_1	A_2A_3	A_3A_3

Solution 1: The log likelihood ratios for each relationship are shown in Table 5.3 for the seven loci combined. The relationship $R_{2:2,1}$ as $\{X_2, X_3\}\{X_1\}$ is inferred as the true relationship when the inference is based on the first rank only.

Solution 2: The simulation experiments were carried out as before and the following results obtained. Each entry in the table represents the number of times each relationship is placed in each rank (or excluded) from 1000 replicate trials.

		Rank				
Relationship	exc	1	2	3	4	5
{iii} simulated						
{iii}	0	699	185	51	45	20
{ii}{i}	0	94	254	310	217	126
{i}{i}{i}	0	21	63	90	230	596
{ii}{j} simulated						
{ij}	88	117	205	171	235	184
{ii}{j}	0	621	198	111	46	24
{ij}{i}	0	68	58	278	297	200
{i}{i}{j}	0	129	481	169	141	80
{i}{j}{k} simulated						
{ijk}	175	33	63	67	234	428
{ij}{k}	0	168	181	288	236	127
{i}{j}{k}	0	463	395	71	56	15

The observed pattern of results for Example C shows some similarity to the expected inference pattern when the true relationship is either $R_{2:2,1}$ or $R_{3:1,1,1}$. If $R_{2:2,1}$ is the true relationship, however, the relationship $R_{1:3}$ would be expected to be ranked higher than fifth position. Given the small difference between the likelihoods for the first two ranked relationships, it seems reasonable to conclude that the best estimate of the relationship between these three individuals is $R_{3:1,1,1}$.

Rank	Relationship		Log Likelihood Ratio
	R_*	as	
1	$R_{2:2,1}$	$\{X_2, X_3\}\{X_1\}$	0.0705
2	$R_{3:1,1,1}$	$\{X_1\}\{X_2\}\{X_3\}$	0.0000
3	$R_{2:2,1}$	$\{X_1, X_3\}\{X_2\}$	-1.8232
4	$R_{2:2,1}$	$\{X_1, X_2\}\{X_3\}$	-2.1586
5	$R_{1:3}$	$\{X_1, X_2, X_3\}$	-4.3570

Table 5.3: The results for Example C when the likelihoods for each relationship are combined over seven loci.

Correct Solution These three individuals were simulated to be full sibs. In contrast to the previous two examples, the inference of relationship for Example C is completely wrong.

Example D

Sample Genotypes: There are five individuals in this sample and genotype data from eight loci; the number and frequency of alleles are given below. Locus 2 provides little information regarding relatedness as all five individuals are homozygous for the common allele.

X	Locus number and p							
	1	2	3	4	5	6	7	8
	(0.2,0.8)	(0.2,0.8)	(0.6,0.4)	(0.8,0.2)	(0.4,0.6)	(0.6,0.4)	(0.2,0.3,0.5)	(0.2,0.3,0.5)
1	A_1A_1	A_2A_2	A_1A_2	A_1A_1	A_2A_2	A_1A_1	A_2A_3	A_2A_3
2	A_1A_2	A_2A_2	A_1A_1	A_1A_1	A_1A_2	A_2A_2	A_1A_3	A_3A_3
3	A_1A_1	A_2A_2	A_1A_1	A_1A_2	A_1A_2	A_1A_2	A_3A_3	A_2A_2
4	A_1A_1	A_2A_2	A_1A_1	A_1A_2	A_2A_2	A_2A_2	A_3A_3	A_1A_3
5	A_1A_1	A_2A_2	A_1A_2	A_1A_2	A_1A_2	A_1A_2	A_1A_3	A_3A_3

When there are five individuals there are 52 possible relationships; with the genotypes observed in Example D, seven of the 52 relationships are excluded (see Table 5.4). It is interesting to note that individuals X_3 and X_4 are found as putative sibs in all the excluded relationships, but they are also found as sibs in the two highest ranked relationships. Relationships are excluded when X_3 and X_4 are considered as siblings of X_2 or X_5 , which indicates that X_3 and X_4 form a family of two sibs.

Rank	Relationship		Log Likelihood
	$R_.$	as	
1	$R_{3:2,2,1}$	$\{X_2, X_5\}\{X_3, X_4\}\{X_1\}$	1.0046
2	$R_{4:2,1,1,1}$	$\{X_3, X_4\}\{X_1\}\{X_2\}\{X_5\}$	0.5268
3	$R_{3:3,1,1}$	$\{X_2, X_4, X_5\}\{X_1\}\{X_3\}$	0.4951
4	$R_{4:2,1,1,1}$	$\{X_2, X_5\}\{X_1\}\{X_3\}\{X_4\}$	0.4777
5	$R_{5:1,1,1,1,1}$	$\{X_1\}\{X_2\}\{X_3\}\{X_4\}\{X_5\}$	0.0000
6	$R_{4:2,1,1,1}$	$\{X_2, X_4\}\{X_1\}\{X_3\}\{X_5\}$	-0.2014
7	$R_{4:2,1,1,1}$	$\{X_4, X_5\}\{X_1\}\{X_2\}\{X_3\}$	-0.4540
8	$R_{4:2,1,1,1}$	$\{X_3, X_5\}\{X_1\}\{X_2\}\{X_4\}$	-0.5017
9	$R_{3:2,2,1}$	$\{X_2, X_4\}\{X_3, X_5\}\{X_1\}$	-0.7031
10	$R_{3:3,1,1}$	$\{X_2, X_3, X_5\}\{X_1\}\{X_4\}$	-0.9832
11	$R_{2:3,2}$	$\{X_2, X_4, X_5\}\{X_1, X_3\}$	-1.0332
12	$R_{3:2,2,1}$	$\{X_1, X_3\}\{X_2, X_5\}\{X_4\}$	-1.0506
13	$R_{3:2,2,1}$	$\{X_1, X_5\}\{X_3, X_4\}\{X_2\}$	-1.1132
14	$R_{2:3,2}$	$\{X_1, X_3, X_4\}\{X_2, X_5\}$	-1.2228
15	$R_{4:2,1,1,1}$	$\{X_1, X_3\}\{X_2\}\{X_4\}\{X_5\}$	-1.5283
16	$R_{4:2,1,1,1}$	$\{X_1, X_5\}\{X_2\}\{X_3\}\{X_4\}$	-1.6401
17	$R_{3:3,1,1}$	$\{X_1, X_3, X_4\}\{X_2\}\{X_5\}$	-1.7005
18	$R_{3:2,2,1}$	$\{X_1, X_3\}\{X_2, X_4\}\{X_5\}$	-1.7297
19	$R_{3:2,2,1}$	$\{X_1, X_4\}\{X_2, X_5\}\{X_3\}$	-1.8238
20	$R_{3:2,2,1}$	$\{X_1, X_5\}\{X_2, X_4\}\{X_3\}$	-1.8415
21	$R_{4:2,1,1,1}$	$\{X_2, X_3\}\{X_1\}\{X_4\}\{X_5\}$	-1.8914
22	$R_{3:2,2,1}$	$\{X_1, X_3\}\{X_4, X_5\}\{X_2\}$	-1.9823
23	$R_{4:2,1,1,1}$	$\{X_1, X_4\}\{X_2\}\{X_3\}\{X_5\}$	-2.3015
24	$R_{3:2,2,1}$	$\{X_2, X_3\}\{X_4, X_5\}\{X_1\}$	-2.3454
25	$R_{2:3,2}$	$\{X_1, X_2, X_5\}\{X_3, X_4\}$	-2.5805
26	$R_{3:2,2,1}$	$\{X_1, X_2\}\{X_3, X_4\}\{X_5\}$	-2.7554
27	$R_{3:2,2,1}$	$\{X_1, X_4\}\{X_3, X_5\}\{X_2\}$	-2.8032
28	$R_{3:3,1,1}$	$\{X_1, X_3, X_5\}\{X_2\}\{X_4\}$	-3.0428
29	$R_{3:3,1,1}$	$\{X_1, X_2, X_5\}\{X_3\}\{X_4\}$	-3.1074
30	$R_{2:3,2}$	$\{X_1, X_3, X_5\}\{X_2, X_4\}$	-3.2442
31	$R_{4:2,1,1,1}$	$\{X_1, X_2\}\{X_3\}\{X_4\}\{X_5\}$	-3.2823
32	$R_{2:3,2}$	$\{X_2, X_3, X_5\}\{X_1, X_4\}$	-3.2847
33	$R_{3:2,2,1}$	$\{X_1, X_5\}\{X_2, X_3\}\{X_4\}$	-3.5314
34	$R_{3:2,2,1}$	$\{X_1, X_2\}\{X_4, X_5\}\{X_3\}$	-3.7363
35	$R_{3:2,2,1}$	$\{X_1, X_2\}\{X_3, X_5\}\{X_4\}$	-3.7840
36	$R_{3:2,2,1}$	$\{X_1, X_4\}\{X_2, X_3\}\{X_5\}$	-4.1928
37	$R_{3:3,1,1}$	$\{X_1, X_4, X_5\}\{X_2\}\{X_3\}$	-5.0735
38	$R_{2:4,1}$	$\{X_1, X_2, X_3, X_5\}\{X_4\}$	-5.1650
39	$R_{2:4,1}$	$\{X_1, X_2, X_4, X_5\}\{X_3\}$	-5.8093
40	$R_{3:3,1,1}$	$\{X_1, X_2, X_3\}\{X_4\}\{X_5\}$	-6.0667
41	$R_{2:3,2}$	$\{X_1, X_2, X_3\}\{X_4, X_5\}$	-6.5207
42	$R_{3:3,1,1}$	$\{X_1, X_2, X_4\}\{X_3\}\{X_5\}$	-6.7111
43	$R_{2:3,2}$	$\{X_1, X_4, X_5\}\{X_2, X_3\}$	-6.9649
44	$R_{2:3,2}$	$\{X_1, X_2, X_4\}\{X_3, X_5\}$	-7.2128
0	$R_{1:5}$	$\{X_1, X_2, X_3, X_4, X_5\}$	Excluded
0	$R_{2:4,1}$	$\{X_1, X_2, X_3, X_4\}\{X_5\}$	Excluded
0	$R_{2:4,1}$	$\{X_1, X_3, X_4, X_5\}\{X_2\}$	Excluded
0	$R_{2:4,1}$	$\{X_2, X_3, X_4, X_5\}\{X_1\}$	Excluded
0	$R_{2:3,2}$	$\{X_2, X_3, X_4\}\{X_1, X_5\}$	Excluded
0	$R_{2:3,2}$	$\{X_3, X_4, X_5\}\{X_1, X_2\}$	Excluded
0	$R_{3:3,1,1}$	$\{X_2, X_3, X_4\}\{X_1\}\{X_5\}$	Excluded
0	$R_{3:3,1,1}$	$\{X_3, X_4, X_5\}\{X_1\}\{X_2\}$	Excluded

Table 5.4: The likelihoods for each of the 52 relationships that could exist between the five individuals comprising Example D. These are the totals for eight loci combined.

Solution 1: Combining the results for all loci, the relationship $R_{3:2,2,1}$ as $\{X_2, X_5\}$ $\{X_3, X_4\}\{X_1\}$ is inferred as the true relationship, based on the first rank only.

Solution 2: The simulation tests were carried out with the appropriate locus parameters and the single sample results compared with the expected results.

The relationships for five individuals can be divided into seven groups according to the number and size of the families: $R_{1:5}$, $R_{2:4,1}$, $R_{2:3,2}$, $R_{3:3,1,1}$, $R_{3:2,2,1}$, $R_{4:2,1,1,1}$ and $R_{5:1,1,1,1,1}$. A representative relationship from each group is simulated for 1000 trials, in the same way as for the previous examples, and the results recorded. The numerical results are not given in detail here; instead three sets of results are presented graphically to illustrate how the inference of relationship is made.

In each figure, the horizontal axis represents the 52 ranks; the number of times each relationship is excluded is given by rank zero. The seven relationship groups, $R_{1:5}$ – $R_{5:1,1,1,1,1}$, are shown from top to bottom of the page. For each group, each relationship is represented by one row in the shaded graph; thus there are 52 rows in total, separated into seven blocks.

The level of shading indicates the number of relationships placed in each rank, the higher the number the darker the square. For example, in Figure 5.1a the simulated relationship is $R_{5:1,1,1,1,1}$. The darker parts of the graph indicate that in the simulations, relationship $R_{5:1,1,1,1,1}$ (bottom row) is placed most often in the higher ranks whereas $R_{1:5}$ (top row) is placed most often in the lowest ranks. This is consistent with results found previously. Note also that the relationships with families of three or more are often excluded: this is seen by the shading for rank zero.

Above each shaded block are points representing the results from Example D; these points are therefore the same in Figures 5.1a to 5.1c. For example, the relationship $R_{1:5}$ is excluded and so the point for this relationship appears above rank zero.

In order to estimate the relationship between the individuals in Example D, the pattern of results from Example D is compared with each of the simulation results in turn. The best estimate of the relationship in Example D is found when the points lie in approximately the same horizontal position as the dark regions of the plots. A visual inspection of the graphs reveals that the points generally correspond more closely

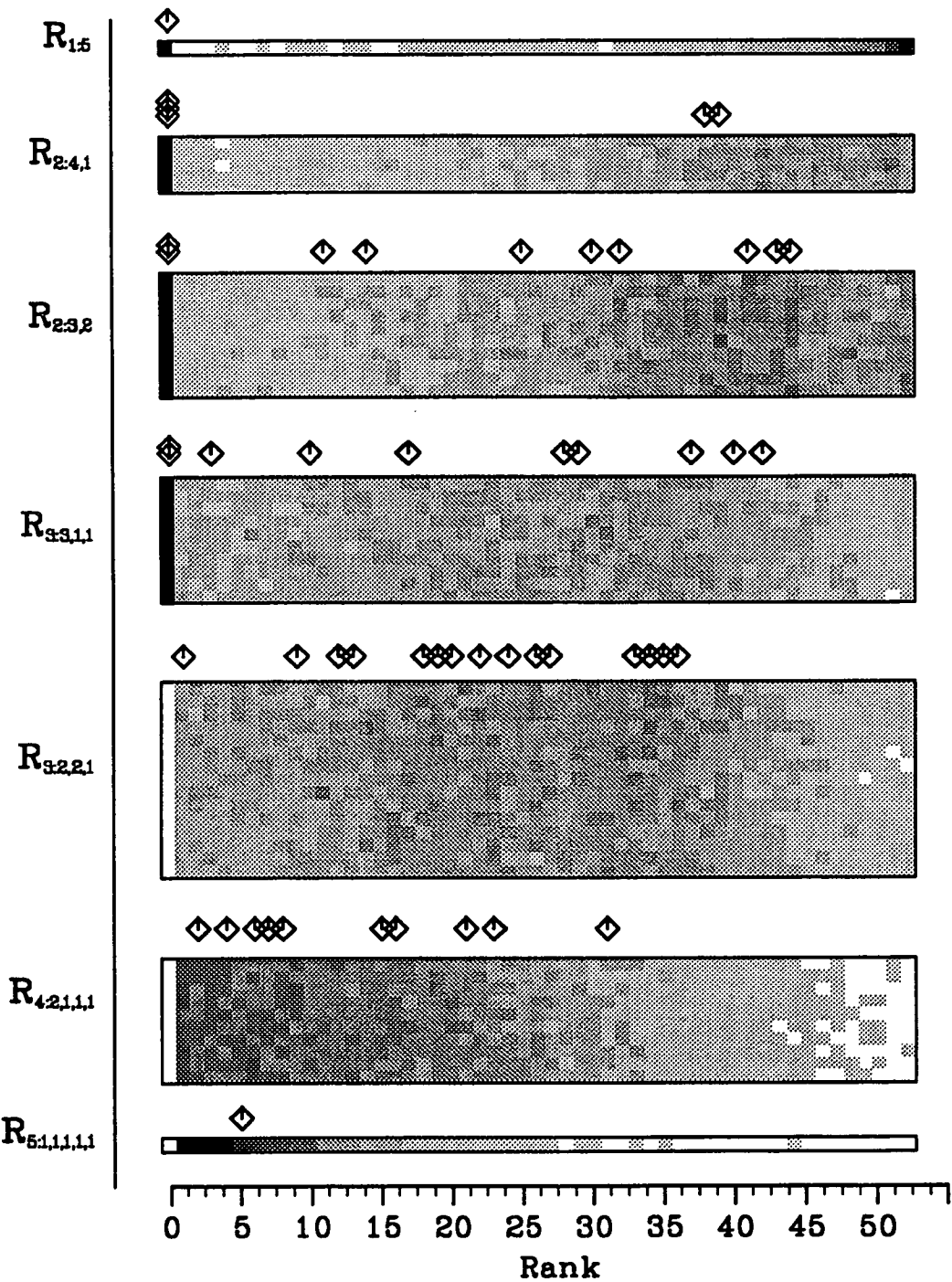


Figure 5.1a: The pattern of results observed from simulation experiments, given that relationship $R_{5,1,1,1,1,1}$ is true. The points above each section show the results observed for the single sample, Example D (see text).

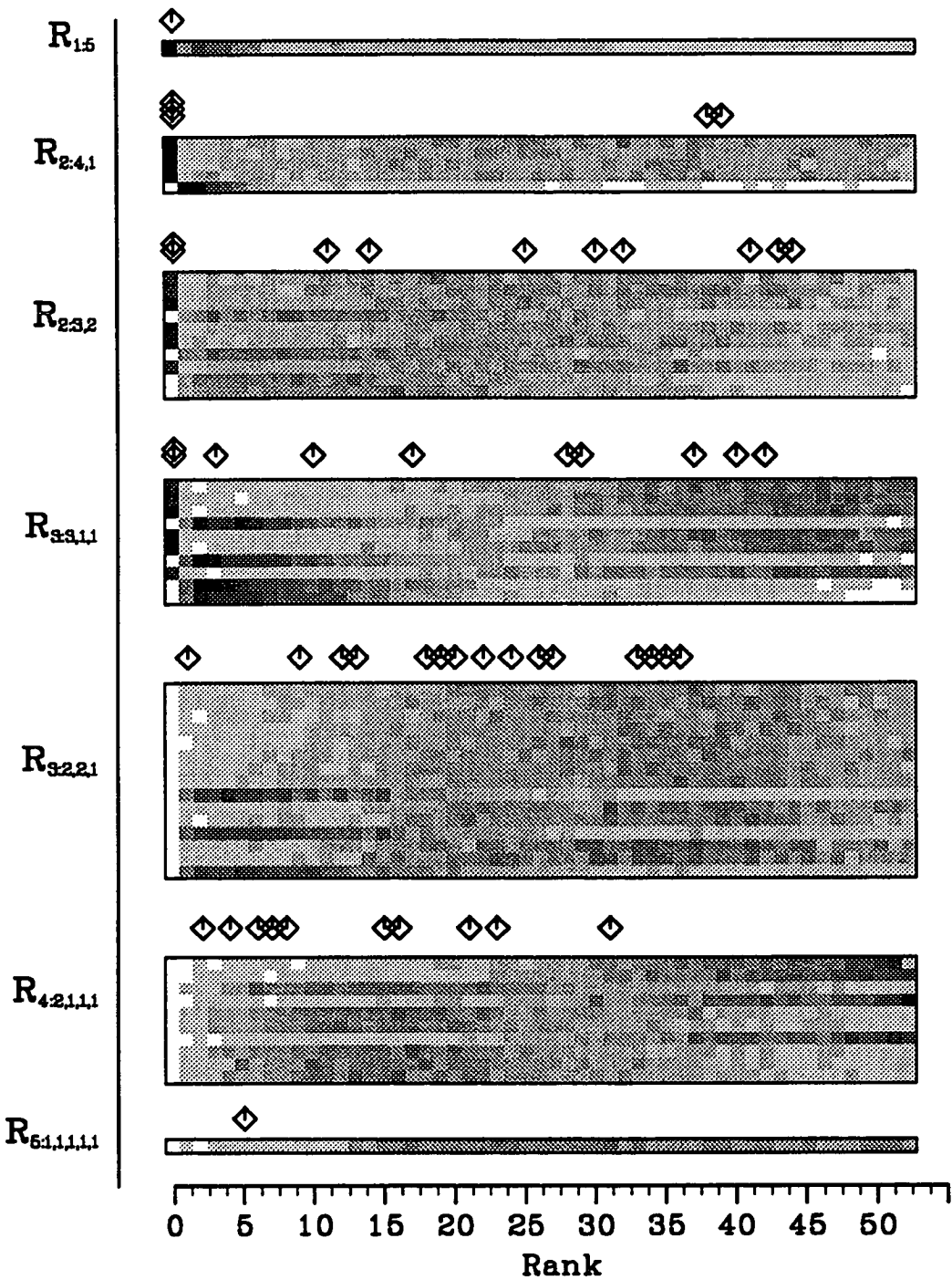


Figure 5.1b: The inference pattern observed from simulation experiments, when the true relationship is $R_{2,4,1}$. The points above each section show the results observed for the single sample, Example D.

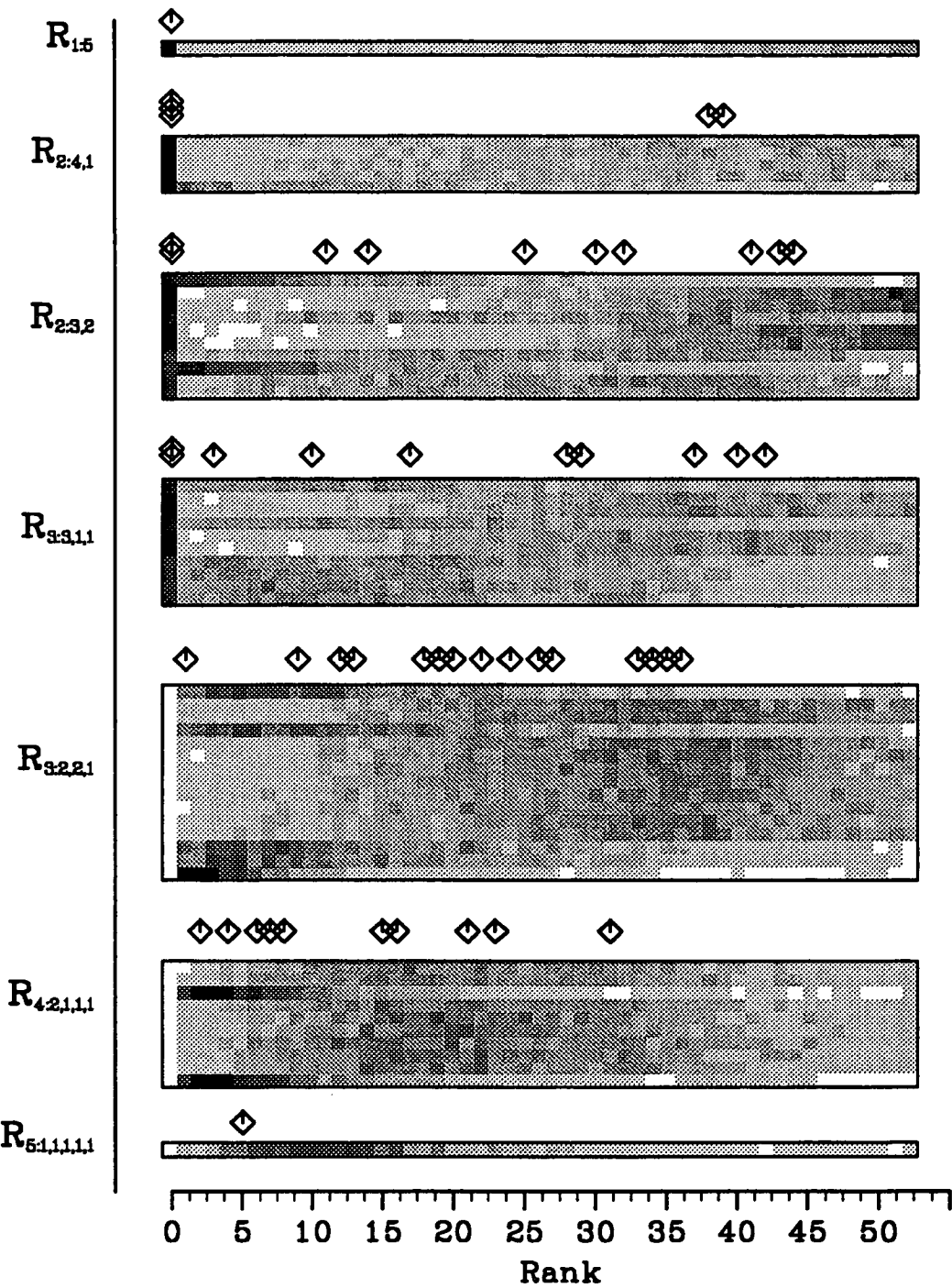


Figure 5.1c: The inference pattern observed from simulation experiments, when the true relationship is $R_{3,2,2,1}$. The points above each section show the results observed for the single sample, Example D (see text).

with the observed patterns for the simulated relationships with smaller sibships, e.g. $R_{3:2,2,1}$, $R_{4:2,1,1,1}$ and $R_{5:1,1,1,1,1}$. Figure 5.1c shows the pattern of inference when the true relationship is $R_{2:4,1}$ and illustrates the poor fit of points to the simulation results. The points correspond closely to the simulation results for both $R_{5:1,1,1,1,1}$ and $R_{3:2,2,1}$ (Figures 5.1c and 5.1b).

Given that the relationship $R_{3:2,2,1}$ as $\{X_2, X_5\}\{X_3, X_4\}\{X_1\}$ is ranked first, this remains the best estimate of the relatedness of the individuals in Example D.

Correct Solution The true relationship between the five individuals in Example D is $R_{5:1,1,1,1,1}$. Although this relationship was not correctly inferred, the pattern of relationships in the ranked likelihoods does indicate that the five individuals are not siblings.

5.3 An Example for a Natural Population

Data for the following example were taken from Quinn *et al* [1987]. Field observations had suggested that lesser snow geese (*Anser caerulescens caerulescens*) engage in both intraspecific brood parasitism (where eggs are laid in nests of adults other than that of the biological parents) and extra-pair fertilisation. These birds were thought to be normally a socially monogamous species. In order to study this, Quinn *et al* looked at 14 polymorphic loci to examine relatedness among goslings in four separate nests. In total, 17 probes were used to identify restriction fragment length polymorphisms (see Section 1.4). Quinn *et al* gave genotypes for the two putative parents and a number of goslings for each of four nests.

For the likelihood analysis presented in this thesis, it is assumed that there is no information from the parental generation. All data from the adult nest attendants is therefore ignored for the analysis, but is used to assess the results.

This set of examples is unlike the ones given previously in that the true population allele frequencies are unknown. The analysis presented in Chapters 2 to 4 was developed on the assumption that the allele frequencies were known. In order to carry out the

DNA marker		Estimated Frequency				DNA marker		Est. Frequency	
		A_1	A_2	A_3	A_4			A_1	A_2
1	DQSG1-AI	0.206	0.647	0.147		8	DQSG7-BI	0.412	0.588
2	DQSG1-AII	0.059	0.941			9	DQSG7-BII	0.441	0.559
3	DQSG1-CI	0.265	0.529	0.147	0.059	10	DQSG7-BIII	0.941	0.059
4	DQSG5-BI	0.794	0.206			11	DQSG8-BI	0.794	0.206
5	DQSG5-BII	0.824	0.176			12	DQSG8-BII	0.824	0.176
6	DQSG5-BIII	1.000				13	DQSG8-DI	0.559	0.441
7	DQSG7-AI	0.735	0.176	0.059	0.029	14	DQSG8-DII	0.471	0.529

DNA marker	Nest 1				Nest 2			
	G1	G2	G3	G4	G5	G6	G7	G8
1	A_2A_2	A_2A_2	A_2A_2	A_2A_3	A_1A_3	A_1A_1	A_1A_3	A_1A_3
2	A_2A_2	A_2A_2	A_2A_2	A_2A_2	A_2A_2	A_2A_2	A_2A_2	A_2A_2
3	A_2A_2	A_2A_2	A_2A_2	A_1A_1	A_2A_2	A_2A_2	A_1A_3	A_1A_2
4	A_1A_2	A_1A_1	A_1A_2	A_1A_1	A_1A_2	A_1A_1	A_1A_1	A_1A_1
5	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_2	A_1A_1	A_1A_2	A_1A_2
6	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1
7	A_1A_1	A_3A_4	A_1A_3	A_1A_2	A_1A_2	A_1A_1	A_1A_1	A_1A_1
8	A_1A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_1	A_1A_2	A_2A_2	A_1A_1
9	A_1A_2	A_2A_2	A_1A_2	A_1A_2	A_1A_2	A_1A_2	A_1A_2	A_1A_2
10	A_1A_1	A_1A_2	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1
11	A_1A_2	A_1A_2	A_1A_1	A_1A_1	A_1A_1	A_1A_2	A_1A_2	A_1A_2
12	A_1A_1	A_1A_2	A_2A_2	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1
13	A_1A_2	A_2A_2	A_2A_2	A_1A_1	A_1A_2	A_1A_2	A_1A_2	A_1A_2
14	A_1A_2	A_2A_2	A_2A_2	A_1A_1	A_1A_2	A_1A_2	A_1A_2	A_1A_2

DNA marker	Nest 3					Nest 4			
	G9	G10	G11	G12	G13	G14	G15	G16	G17
1	A_2A_2	A_2A_2	A_1A_2	A_2A_3	A_1A_2	A_2A_2	A_2A_2	A_2A_2	A_2A_2
2	A_2A_2	A_2A_2	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2	A_2A_2
3	A_1A_2	A_2A_4	A_2A_4	A_2A_2	A_2A_2	A_1A_3	A_1A_3	A_1A_3	A_1A_3
4	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_2	A_1A_2	A_1A_2	A_1A_2
5	A_1A_2	A_1A_2	A_1A_1	A_1A_1	A_1A_2	A_1A_1	A_1A_1	A_1A_1	A_1A_1
6	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1
7	A_1A_1	A_1A_2	A_1A_1	A_2A_2	A_1A_2	A_1A_1	A_1A_1	A_1A_1	A_1A_1
8	A_2A_2	A_1A_1	A_1A_2	A_1A_1	A_2A_2	A_1A_2	A_1A_2	A_2A_2	A_2A_2
9	A_1A_2	A_1A_2	A_2A_2	A_1A_2	A_1A_2	A_1A_2	A_1A_2	A_1A_2	A_1A_2
10	A_1A_2	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1
11	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_2	A_1A_1	A_1A_2	A_1A_1
12	A_1A_1	A_1A_1	A_1A_1	A_1A_1	A_1A_2	A_1A_1	A_1A_2	A_1A_1	A_1A_2
13	A_1A_2	A_2A_2	A_1A_1	A_1A_1	A_1A_2	A_1A_2	A_1A_1	A_1A_2	A_1A_1
14	A_1A_2	A_1A_1	A_1A_1	A_2A_2	A_1A_2	A_1A_2	A_1A_2	A_1A_2	A_2A_2

Table 5.5: The 14 loci with estimated allele frequencies, and the genotypes for the 17 goslings taken from four lesser snow goose nests.

analysis in this chapter, the frequencies are estimated from the data as a simple count of alleles from the 17 goslings in the study.

Let the alleles be $A_1, A_2, \dots$ for each marker; the estimated allele frequencies for each allele are shown in Table 5.5. Two of the markers have four alleles, and one has three alleles. In all cases one of the alleles is very common. The remaining markers have only two alleles except DQSG5-BIII (locus 6) which is invariant. The observed genotypes for the 17 goslings used in the experimental population are also shown in Table 5.5.

Nest Number 1

This nest has four goslings and these are labelled $G1-G4$ to correspond with the published notation. Using the allele frequencies estimated from the 17 goslings, the likelihoods for each of the 15 possible relationships are calculated and ranked in descending order (see Table 5.6). The relationships $R_{1.5} (\{G1, G2, G3, G4\})$ and $R_{2.3,1}$ as $\{G1, G2, G4\}\{G3\}$ are excluded, showing that the four goslings cannot all be full sibs based on the genotypes recorded for the goslings.

Solution 1: From the observed genotype data for nest 1, it is inferred that relationship $R_{2.3,1}$ as $\{G1, G2, G3\}\{G4\}$ is the best estimate of the relationship between the goslings. Thus gosling $G4$ is not the full sibling of the other three goslings in the nest.

Solution 2: The two relationships placed first and second have high likelihood values; both have gosling $G4$ as unrelated so this suggests that gosling number 4 is not the offspring of the same two parents as the other three goslings. Using the allele frequencies for the 14 loci to generate random samples of four individuals, expected results were found as described previously. The general pattern seen in the ordering of the relationships is similar to that found in the previously in Section 5.2 (Examples A and B), but tend to be more extreme, with the correct relationship being inferred more often. For example, 791 samples are correctly inferred as $R_{1.4}$ whereas only 578 such samples are correctly inferred using simulated data from the eight loci of Example A.

Rank	Relationship		Log Likelihood Ratio
	R_{*}	as	
1	$R_{2:3,1}$	$\{G1, G2, G3\}\{G4\}$	4.0375
2	$R_{3:2,1,1}$	$\{G2, G3\}\{G1\}\{G4\}$	3.0127
3	$R_{4:1,1,1,1}$	$\{G1\}\{G2\}\{G3\}\{G4\}$	0.0000
4	$R_{2:2,2}$	$\{G1, G4\}\{G2, G3\}$	-0.0091
5	$R_{3:2,1,1}$	$\{G1, G3\}\{G2\}\{G4\}$	-1.0494
6	$R_{2:3,1}$	$\{G2, G3, G4\}\{G1\}$	-2.0718
7	$R_{3:2,1,1}$	$\{G1, G2\}\{G3\}\{G4\}$	-2.4327
8	$R_{3:2,1,1}$	$\{G1, G4\}\{G2\}\{G3\}$	-3.1040
9	$R_{3:2,1,1}$	$\{G3, G4\}\{G1\}\{G2\}$	-7.0237
10	$R_{3:2,1,1}$	$\{G2, G4\}\{G1\}\{G3\}$	-8.7602
11	$R_{2:2,2}$	$\{G1, G2\}\{G3, G4\}$	-9.4564
12	$R_{2:2,2}$	$\{G1, G3\}\{G2, G4\}$	-9.8096
13	$R_{2:3,1}$	$\{G1, G3, G4\}\{G2\}$	-9.8693
0	$R_{1:4}$	$\{G1, G2, G3, G4\}$	Excluded
0	$R_{2:3,1}$	$\{G1, G2, G4\}\{G3\}$	Excluded

Table 5.6: The log likelihood ratios for each relationship for goslings $G1, G2, G3$ and $G4$ from nest 1.

This may be due partly to the increased amount of data (14 loci) and partly to the extreme allele frequencies calculated from the small snow goose population.

The pattern of relationships in Table 5.6 corresponds closely with the patterns observed when $R_{3:2,1,1}$ as $\{G2, G3\}\{G1\}\{G4\}$ or $R_{2:2,2}$ as $\{G1, G4\}\{G2, G3\}$ are the simulated relationships. This suggests that goslings $G2$ and $G3$ are full sibs but goslings $G1$ and $G4$ do not have the same two parents as $G2$ and $G3$. The likelihood analysis only considers 15 hypothetical relationships in which the individuals are either full sibs or unrelated. In this natural population, however, gosling $G1$ may be a half sib of $G2$ and $G3$.

Comparison of Results: Direct comparison of the goslings with the genotypes of the nest attendants [Quinn *et al*, 1987] reveals that gosling $G4$ can indeed be excluded as an offspring of the nest attendants as a pair, because it has alleles that are not present in either adult. Gosling $G1$ cannot positively be excluded as an offspring of the nest attendants on the basis of the genotypes, but may be a half sib.

Rank	Relationship		Log Likelihood
	$R_{\cdot}$	as	
1	$R_{1:4}$	$\{G5, G6, G7, G8\}$	6.0871
2	$R_{2:3,1}$	$\{G6, G7, G8\}\{G5\}$	5.7053
3	$R_{2:3,1}$	$\{G5, G7, G8\}\{G6\}$	4.7828
4	$R_{2:3,1}$	$\{G5, G6, G8\}\{G7\}$	3.7446
5	$R_{3:2,1,1}$	$\{G7, G8\}\{G5\}\{G6\}$	2.9036
6	$R_{2:2,2}$	$\{G5, G8\}\{G6, G7\}$	2.7297
7	$R_{2:2,2}$	$\{G5, G6\}\{G7, G8\}$	2.6260
8	$R_{3:2,1,1}$	$\{G5, G8\}\{G6\}\{G7\}$	2.5474
9	$R_{3:2,1,1}$	$\{G6, G8\}\{G5\}\{G7\}$	1.4819
10	$R_{2:2,2}$	$\{G5, G7\}\{G6, G8\}$	0.5045
11	$R_{3:2,1,1}$	$\{G6, G7\}\{G5\}\{G8\}$	0.1823
12	$R_{2:3,1}$	$\{G5, G6, G7\}\{G8\}$	0.1102
13	$R_{4:1,1,1,1}$	$\{G5\}\{G6\}\{G7\}\{G8\}$	0.0000
14	$R_{3:2,1,1}$	$\{G5, G6\}\{G7\}\{G8\}$	-0.2776
15	$R_{3:2,1,1}$	$\{G5, G7\}\{G6\}\{G8\}$	-0.9775

Table 5.7: The log likelihood ratios for each relationship for goslings $G5$ – $G8$ from nest 2.

Nest Number 2

The second nest also has four gosling, labelled $G5$ – $G8$. The results from the likelihood analysis are shown in Table 5.7. None of the 15 possible relationships are excluded.

Solution 1: The relationship $R_{1:4}$ is ranked first and hence these four gosling are inferred to be full siblings.

Solution 2: In contrast to nest 1, these four goslings appear to be closely related, the relationship $\{G5, G6, G7, G8\}$ maximises the probability of the data and relationships with several families are ranked lower. Comparing the pattern of relationships in Table 5.7 with the simulation results, the best estimate for the relationship between the goslings in nest 2 is that they are full sibs.

Comparison of Results: This nest was specifically chosen because both nest attendants were white whereas gosling $G5$ was blue. Colouring is known to be determined

by a single autosomal gene, with white being a recessive character. This gives direct evidence for gosling *G5* being the result of extra-pair fertilisation or intraspecific brood parasitism. However, looking at only the genotypes of the offspring at the chosen loci, the pattern of ranked relationships does not suggest that gosling *G5* is unusual.

Nest Number 3

As there are five individuals in this nest, there are 52 possible relationships. The likelihoods for each relationship are given in Table 5.8.

Solution 1: Combining the information from all 14 loci, $R_{4:2,1,1,1}$ as $\{G10, G11\} \{G9\} \{G12\} \{G13\}$ emerges as the highest ranked relationship.

Solution 2: Like nest 1, the individuals in this nest show a pattern of relationship that indicates the goslings are not all full sibs. None of the 52 structures are positively excluded but there is a general trend in the ranking such that relationships with few, large sibships are ranked lower. The top five relationships have similar likelihoods and gosling *G12* is placed separately in all these higher ranked structures. Examining the pattern of relationships in the 52 ranks, and comparing the relative ordering with the simulation results, some similarity is found in the pattern for both $R_{4:2,1,1,1}$ and $R_{5:1,1,1,1,1}$. The pattern of results from the simulation experiments (for $R_{5:1,1,1,1,1}$) and the observed results from nest 3 are shown in Figure 5.2; the single sample results match this pattern closely. The best estimate for the relationship between the five individuals in nest 3 is that they are not closely related. Although the goslings are in the same nest, they do not all have the same biological parents. In a wild population like the snow geese, however, other relationships would obviously exist in addition to the 52 described; half siblings cannot be distinguished here.

Comparison of Results: Like nest 2, this particular nest was chosen because both attendants were white while one gosling (*G9*) was blue and therefore cannot be the offspring of both nest attendants. Direct comparison with the genotypes of the nest attendants shows that goslings *G11* and *G12* cannot be the offspring of either attendant and goslings *G9* and *G10* can be the offspring of only one of the adults. Only *G13* is compatible as an offspring of both nest attendants. Quinn *et al* [1987] suggest that one reason for this could be that the attendants took over an abandoned nest. Even if this

Rank	Relationship		Log Likelihood
	<i>R.</i>	as	
1	$R_{4:2,1,1,1}$	$\{G_{10}, G_{11}\}\{G_9\}\{G_{12}\}\{G_{13}\}$	0.1218
2	$R_{3:2,2,1}$	$\{G_9, G_{13}\}\{G_{10}, G_{11}\}\{G_{12}\}$	0.0885
3	$R_{5:1,1,1,1,1}$	$\{G_9\}\{G_{10}\}\{G_{11}\}\{G_{12}\}\{G_{13}\}$	0.0000
4	$R_{4:2,1,1,1}$	$\{G_9, G_{13}\}\{G_{10}\}\{G_{11}\}\{G_{12}\}$	-0.0333
5	$R_{3:3,1,1}$	$\{G_{10}, G_{11}, G_{13}\}\{G_9\}\{G_{12}\}$	-0.3735
6	$R_{4:2,1,1,1}$	$\{G_{10}, G_{13}\}\{G_9\}\{G_{11}\}\{G_{12}\}$	-0.9385
7	$R_{4:2,1,1,1}$	$\{G_{10}, G_{12}\}\{G_9\}\{G_{11}\}\{G_{13}\}$	-1.5022
8	$R_{3:2,2,1}$	$\{G_9, G_{13}\}\{G_{10}, G_{12}\}\{G_{11}\}$	-1.5354
9	$R_{4:2,1,1,1}$	$\{G_9, G_{10}\}\{G_{11}\}\{G_{12}\}\{G_{13}\}$	-1.5973
10	$R_{3:2,2,1}$	$\{G_{10}, G_{11}\}\{G_{12}, G_{13}\}\{G_9\}$	-1.8519
11	$R_{4:2,1,1,1}$	$\{G_{12}, G_{13}\}\{G_9\}\{G_{10}\}\{G_{11}\}$	-1.9736
12	$R_{4:2,1,1,1}$	$\{G_{11}, G_{13}\}\{G_9\}\{G_{10}\}\{G_{12}\}$	-2.0988
13	$R_{3:3,1,1}$	$\{G_9, G_{10}, G_{11}\}\{G_{12}\}\{G_{13}\}$	-2.1754
14	$R_{4:2,1,1,1}$	$\{G_9, G_{11}\}\{G_{10}\}\{G_{12}\}\{G_{13}\}$	-2.7043
15	$R_{4:2,1,1,1}$	$\{G_{11}, G_{12}\}\{G_9\}\{G_{10}\}\{G_{13}\}$	-2.7442
16	$R_{3:2,2,1}$	$\{G_9, G_{13}\}\{G_{11}, G_{12}\}\{G_{10}\}$	-2.7775
17	$R_{3:3,1,1}$	$\{G_{10}, G_{11}, G_{12}\}\{G_9\}\{G_{13}\}$	-3.2181
18	$R_{2:3,2}$	$\{G_{10}, G_{11}, G_{12}\}\{G_9, G_{13}\}$	-3.2514
19	$R_{3:3,1,1}$	$\{G_9, G_{10}, G_{13}\}\{G_{11}\}\{G_{12}\}$	-3.4546
20	$R_{3:2,2,1}$	$\{G_9, G_{10}\}\{G_{12}, G_{13}\}\{G_{11}\}$	-3.5709
21	$R_{3:2,2,1}$	$\{G_{10}, G_{12}\}\{G_{11}, G_{13}\}\{G_9\}$	-3.6009
22	$R_{3:2,2,1}$	$\{G_9, G_{11}\}\{G_{10}, G_{13}\}\{G_{12}\}$	-3.6428
23	$R_{3:2,2,1}$	$\{G_{10}, G_{13}\}\{G_{11}, G_{12}\}\{G_9\}$	-3.6827
24	$R_{3:2,2,1}$	$\{G_9, G_{10}\}\{G_{11}, G_{13}\}\{G_{12}\}$	-3.6960
25	$R_{2:4,1}$	$\{G_{10}, G_{11}, G_{12}, G_{13}\}\{G_9\}$	-3.9913
26	$R_{2:4,1}$	$\{G_9, G_{10}, G_{11}, G_{13}\}\{G_{12}\}$	-4.0053
27	$R_{3:3,1,1}$	$\{G_{10}, G_{12}, G_{13}\}\{G_9\}\{G_{11}\}$	-4.0607
28	$R_{2:3,2}$	$\{G_9, G_{10}, G_{11}\}\{G_{12}, G_{13}\}$	-4.1490
29	$R_{3:2,2,1}$	$\{G_9, G_{11}\}\{G_{10}, G_{12}\}\{G_{13}\}$	-4.2065
30	$R_{3:2,2,1}$	$\{G_9, G_{10}\}\{G_{11}, G_{12}\}\{G_{13}\}$	-4.3414
31	$R_{3:2,2,1}$	$\{G_9, G_{12}\}\{G_{10}, G_{11}\}\{G_{13}\}$	-4.3802
32	$R_{4:2,1,1,1}$	$\{G_9, G_{12}\}\{G_{10}\}\{G_{11}\}\{G_{13}\}$	-4.5020
33	$R_{3:3,1,1}$	$\{G_{11}, G_{12}, G_{13}\}\{G_9\}\{G_{10}\}$	-4.6119
34	$R_{3:2,2,1}$	$\{G_9, G_{11}\}\{G_{12}, G_{13}\}\{G_{10}\}$	-4.6779
35	$R_{3:3,1,1}$	$\{G_9, G_{11}, G_{13}\}\{G_{10}\}\{G_{12}\}$	-4.7617
36	$R_{2:3,2}$	$\{G_{10}, G_{11}, G_{13}\}\{G_9, G_{12}\}$	-4.8755
37	$R_{3:2,2,1}$	$\{G_9, G_{12}\}\{G_{10}, G_{13}\}\{G_{11}\}$	-5.4405
38	$R_{2:3,2}$	$\{G_9, G_{12}, G_{13}\}\{G_{10}, G_{11}\}$	-5.8259
39	$R_{3:3,1,1}$	$\{G_9, G_{12}, G_{13}\}\{G_{10}\}\{G_{11}\}$	-5.9477
40	$R_{3:3,1,1}$	$\{G_9, G_{10}, G_{12}\}\{G_{11}\}\{G_{13}\}$	-5.9581
41	$R_{2:3,2}$	$\{G_9, G_{10}, G_{13}\}\{G_{11}, G_{12}\}$	-6.1988
42	$R_{2:3,2}$	$\{G_{11}, G_{12}, G_{13}\}\{G_9, G_{10}\}$	-6.2092
43	$R_{2:3,2}$	$\{G_9, G_{11}, G_{13}\}\{G_{10}, G_{12}\}$	-6.2639
44	$R_{3:2,2,1}$	$\{G_9, G_{12}\}\{G_{11}, G_{13}\}\{G_{10}\}$	-6.6008
45	$R_{2:3,2}$	$\{G_{10}, G_{12}, G_{13}\}\{G_9, G_{11}\}$	-6.7650
46	$R_{2:3,2}$	$\{G_9, G_{10}, G_{12}\}\{G_{11}, G_{13}\}$	-8.0569
47	$R_{2:4,1}$	$\{G_9, G_{10}, G_{11}, G_{12}\}\{G_{13}\}$	-8.7440
48	$R_{2:4,1}$	$\{G_9, G_{10}, G_{12}, G_{13}\}\{G_{11}\}$	-8.8781
49	$R_{1:5}$	$\{G_9, G_{10}, G_{11}, G_{12}, G_{13}\}$	-10.037
50	$R_{3:3,1,1}$	$\{G_9, G_{11}, G_{12}\}\{G_{10}\}\{G_{13}\}$	-10.450
51	$R_{2:4,1}$	$\{G_9, G_{11}, G_{12}, G_{13}\}\{G_{10}\}$	-11.160
52	$R_{2:3,2}$	$\{G_9, G_{11}, G_{12}\}\{G_{10}, G_{13}\}$	-11.388

Table 5.8: The log likelihood ratios for each relationship for goslings *G*₉–*G*₁₃ from nest 3.

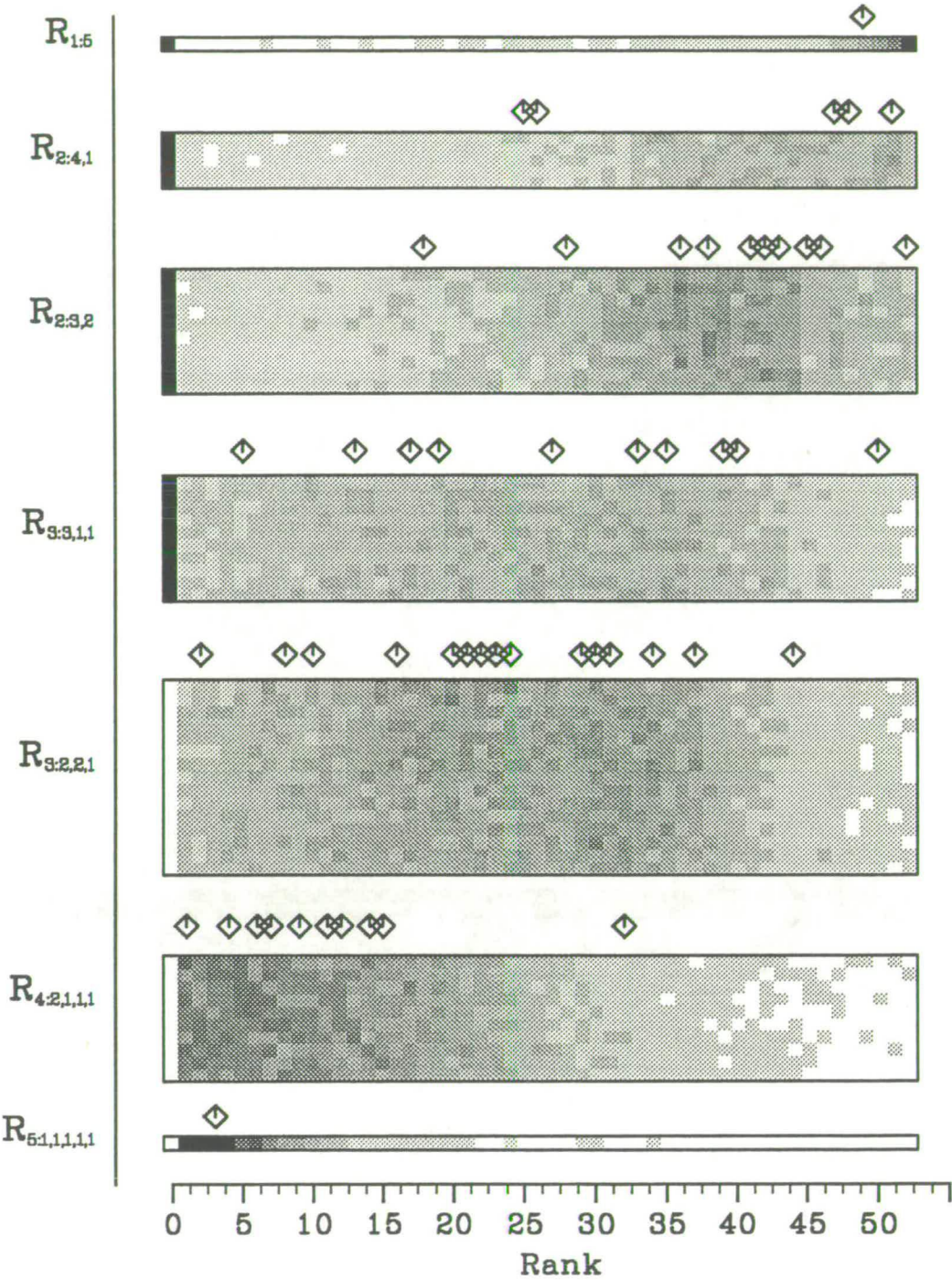


Figure 5.2: The pattern of inference observed from simulation experiments; the true relationship is $R_{5,1,1,1,1,1}$. Points represent the results from the single sample of goslings from nest 3 (See Figure 5.1a and accompanying text for description of graph).

Rank	Relationship		Log Likelihood Ratio
	R_*	as	
1	$R_{1:4}$	$\{G14, G15, G16, G17\}$	11.5968
2	$R_{2:3,1}$	$\{G14, G15, G16\}\{G17\}$	7.3200
3	$R_{2:2,2}$	$\{G14, G16\}\{G15, G17\}$	7.2995
4	$R_{2:3,1}$	$\{G14, G15, G17\}\{G16\}$	6.9219
5	$R_{2:2,2}$	$\{G14, G15\}\{G16, G17\}$	6.2302
6	$R_{2:3,1}$	$\{G15, G16, G17\}\{G14\}$	5.0289
7	$R_{2:3,1}$	$\{G14, G16, G17\}\{G15\}$	4.5405
8	$R_{3:2,2,1}$	$\{G14, G15\}\{G16\}\{G17\}$	3.9629
9	$R_{3:2,2,1}$	$\{G15, G17\}\{G14\}\{G16\}$	3.6623
10	$R_{3:2,2,1}$	$\{G14, G16\}\{G15\}\{G17\}$	3.6372
11	$R_{3:2,1,1}$	$\{G16, G17\}\{G14\}\{G15\}$	2.2674
12	$R_{2:2,2}$	$\{G14, G17\}\{G15, G16\}$	1.6505
13	$R_{3:2,1,1}$	$\{G15, G16\}\{G14\}\{G17\}$	1.1006
14	$R_{3:2,1,1}$	$\{G14, G17\}\{G15\}\{G16\}$	0.5499
15	$R_{4:1,1,1,1}$	$\{G14\}\{G15\}\{G16\}\{G17\}$	0.0000

Table 5.9: The log likelihood ratios for each relationship for goslings G14–G17 from nest 4.

was the case, results from analysis of the goslings do suggest that the eggs present in the nest were not laid by the same pair of (now absent) adults. Although the goslings are not all compatible as offspring of both nest attendants, neither do they seem to be full sibs of each other.

Nest Number 4

Nest 4 has four occupants; the likelihoods for each of the 15 possible relationships are shown in Table 5.9.

Solution 1: Taking all 14 loci simultaneously, the relationship $R_{1:4}$ is inferred as the true relationship.

Solution 2: In direct contrast to nest 3, this example shows the ranking pattern expected for a family of four full sibs. The relationship $\{G14, G15, G16, G17\}$ is ranked first by a very large margin, with other relationships with large sibships occurring high in the rankings. When compared with the set of inference patterns derived from

simulations, this most closely matches that for $R_{1.4}$. These results indicate strongly that nest 4 does contain four full sibs.

Comparison of Results: Quinn *et al* [1987] similarly find no evidence to suggest that the two nest attendants were not the true parents.

5.4 Summary

In this chapter, the likelihood method was applied to single samples of three, four or five individuals. The aim was to draw some useful conclusions regarding the genealogical relationship between the individuals, based on the observed genotypes.

Two types of sample were analysed. Four examples were simulated with known relationship by selecting parental genotypes at random from a large population within which the gene frequencies are known. Four more samples were taken from a natural population so that the true relationship was not known. In the second example, allele frequencies were estimated from the data as the analysis was developed based on known frequencies. The likelihood method was developed specifically for use in such situations where there are small samples and where no information on genotype is available from the parental generation.

For Examples A and B, the true relationship was ranked first and so was correctly inferred; the pattern of relationships in the list of ranked likelihoods confirmed this conclusion. There were only three individuals in Example C, and many of the loci had only two alleles. For this example the wrong relationship was inferred. Example D was a more complex example as there were five individuals and therefore 52 relationship structures considered under this model. Eight of these relationships were excluded. The pattern seen within the list of 44 valid relationships suggested the individuals were not closely related; this conclusion was correct.

For the second set of examples, conclusions regarding relatedness were compared with those of the original authors. The results obtained from analysis of the goslings alone agreed well with those obtained by direct comparison with individuals from the parental generation. It is not possible, however, to distinguish between half sibs in the analysis of the goslings and so the estimate of relationship is essentially a guide to the

degree of relatedness in the sample. The snow goose example highlights the fact that the information gained from analysis using the likelihood method may be useful when used with other methods.

Note

To assess the reliability of the method for inferring relationships, the difference between the likelihoods for a given sample is of interest. In particular, it could be important to know whether the first ranked relationship gives a significantly higher probability for the data than the lower ranked structures. If one hypothesis is a subset of another, then twice the difference in log likelihood is distributed as χ^2 and this can be used as the basis of a significance test. However, the relationship hypotheses proposed here are not nested and thus this method cannot be applied.

Chapter 6

Analysis of Variance Method

6.1 Introduction

The approach described in the previous chapters was developed to investigate the likelihood for genetic relationships among members of a sample of individuals from a population. Emphasis was placed on estimating the detailed relationship between these individuals; a small number of named individuals were considered simultaneously.

There will be occasions, however, when it is neither practical nor desirable to identify precise relationships between named individuals. If the sample for analysis is large, the total number of relationship structures defined by the likelihood model makes exhaustive evaluation too time-consuming. Details of the exact genealogical relationship are not always required; for experiments with plants for example, or field experiments with insects, it is more useful to be able to establish the general pattern of relatedness within a group as opposed to the exact position of every individual in a genealogy. For some investigations it may be sufficient to be able to subdivide the group into smaller units each consisting of individuals that are more closely related to each other than to members of other groups.

In this chapter, in contrast to the likelihood method, the analysis of variance is used to estimate family size for samples drawn from different subdivisions of a population; specific individuals are not identified in this approach. The example used here for illustration is a population of *Drosophila* in which a number of flies colonise a new site and become the founders for the next generation. The cactophilic fly *Drosophila*

buzzatii feeds on the microorganisms associated with decaying prickly pear (*Optunia*) tissues and breeds on the rotting plant. This is an ephemeral and patchy resource and the number of parents breeding on a single plant is limited; individuals associated with a particular plant are usually members of a single generation. It has been shown that heterogeneity at the level of individual breeding rots is important in maintaining polymorphism [Barker *et al*, 1986; Barker, 1992]. An understanding of the breeding structure is therefore important as the basis for further study. Many other *Drosophila* oviposition substrates are sufficiently ephemeral (e.g. small fruits, mushrooms [Jaenike & Selander, 1979; Lacy, 1983]) to support only one generation and are therefore suitable for the analysis described in this chapter.

The question addressed here is: given genotypic data for the first generation of offspring, is it possible to estimate the number of parents that colonised the site?

For this, the second analysis presented in the thesis (named the analysis of variance method), it is assumed that a large number of diploid individuals are sampled from a number of subpopulations in which the members are from the same generation, for example plants or insects in a given region. The data on which inferences of relatedness are to be based are the genotypes of these individuals for one or more autosomal loci (see Section 1.4). An indicator variable, x , is defined to represent the data in the form of a score and an analysis of variance of the scores is then carried out. A comparison of the between and within population subdivision variances is used to estimate the mean family size within the progeny generation; from this it is possible to estimate the number of parents that contributed to the offspring generation.

An analysis of variance was carried out utilising three types of data: allele counts, genotype counts and scores based on heterozygosity. This was done so that the three methods could be compared and the most informative data type chosen; the analysis based on this data type was then applied to real data (Section 6.8). Full details of the calculations are given in Section 6.3 for the analysis based on allele counts, and in less detail in Section 6.4 for the analysis based on genotype counts. The results obtained using heterozygosity data are stated without proof in Section 6.5, as the method is similar to the previous two examples.

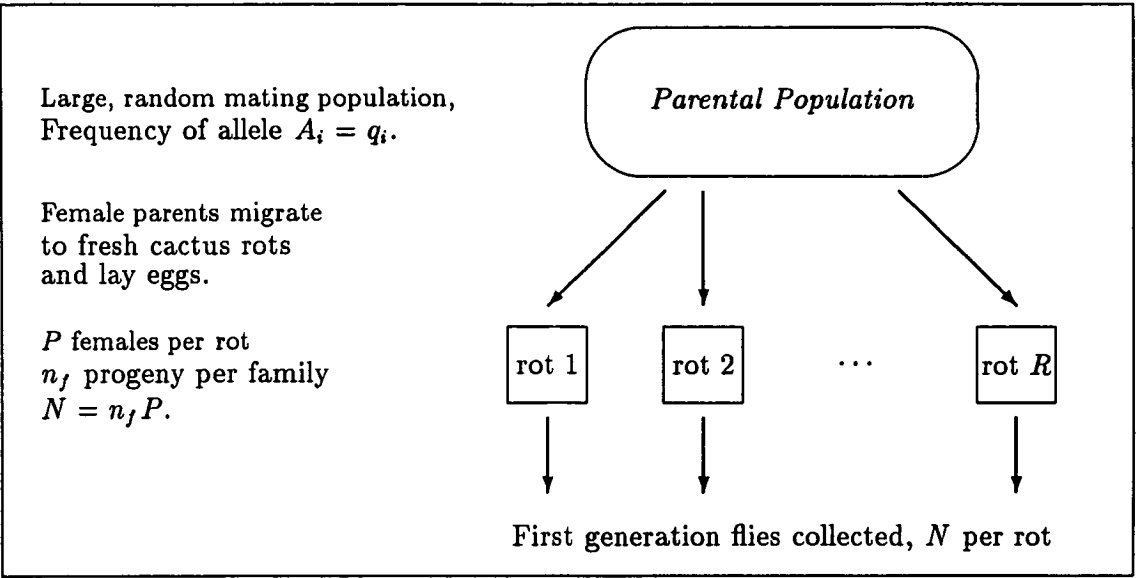


Figure 6.1: The model: a number of adult flies colonise a new site of fresh cactus rots.

6.2 The Model

Natural populations will often be subdivided into smaller groups depending on the habitat and specific behaviour of the organism concerned. For the analyses developed in this thesis, only those populations in which the parent individuals are not directly available for study are considered. One example of such a population is a cactophilic *Drosophila* species; adults colonise new sites for breeding and are not present when the larvae emerge. A rotting cactus cladode (rot) has a limited life as a breeding substrate suitable for *Drosophila* in the field, so fresh rots that appear are colonised by adult flies in the area. The first generation progeny can be collected as they emerge and their genotypes found, thus providing sufficient data for analysis.

The analysis in this chapter is based on a simple, idealised population. The parental generation is a large, natural population of *Drosophila* that colonises a new site of fresh cactus rots. It is assumed that there are no forces acting, such as selection or mutation, and that the population gene frequencies are known. Each female mates at random with one male and lays eggs in one of R rots; the family size (the number of offspring hatched per female) is assumed to be constant and denoted n_f ; a total of P females

may lay eggs at any one rot. In reality, the flies emerging from a rotting cladode tend to remain and mate rather than disperse. Hence it is crucial to know whether the rot has remained viable for two or more generations. For the model, it is assumed that the rot is fresh so that emerging flies can be assumed to be first generation progeny. When the progeny emerge, N flies are collected from each rot where $N = Pn_f$ (see Figure 6.1). The flies are collected as they emerge so that there is no migration between rots.

This is only one representation of the problem; the analysis could be extended to include other more realistic models. For example, the model could be altered to allow for variable numbers of parents per rot and variable family size due to variations in viability and fertility. A more realistic model would also include a degree of inbreeding due to more than one generation surviving per rot.

6.3 Analysis of Variance Based on Allele Scores

For a gene j from an individual i of rot k , let an indicator variable x_{kij} be defined by

$$x_{kij} = \begin{cases} 1 & \text{if the gene is } A_1 \\ 0 & \text{otherwise,} \end{cases}$$

where $k = 1, 2, \dots, R$, $i = 1, 2, \dots, N$ and $j = 1, 2$ for the maternally or paternally inherited alleles. The paternally and maternally inherited alleles within cannot be distinguished, but only autosomal loci are considered and so the alleles are inherited in the same manner. Let each of the RN individuals in the sample have a known genotype for a single autosomal locus that has two alleles. Thus each rot has a total of $2N$ scores associated with it. The expectation of the mean of the variables x_{kij} is q , the population frequency of A_1 .

For an analysis of the variance of the variables x_{kij} see Table 6.1. To estimate the family size within the sample, expressions are obtained for the mean squares within and between rots in terms of n_f . The procedure used to obtain the estimate of family size, $\hat{n}_f$, is given in the following section.

Source of Variation	Degrees of Freedom	Sum of Squares
between rots	$R - 1$	$2N \sum_{k=1}^R (\bar{x}_{k..} - \bar{x}_{...})^2$
within rots	$R(2N - 1)$	$\sum_{k=1}^R \sum_{i=1}^N \sum_{j=1}^2 (x_{kij} - \bar{x}_{k..})^2$
Total	$2RN - 1$	$\sum_{k=1}^R \sum_{i=1}^N \sum_{j=1}^2 (x_{kij} - \bar{x}_{...})^2$

Table 6.1: The analysis of variance table using data based on allele scores.

6.3.1 The Expected Mean Square Between Rots

From Table 6.1, the expected mean squared deviation from the mean between rots is

$$E[MS_b] = \frac{2N}{R-1} E \left[\sum_{k=1}^R (\bar{x}_{k..} - \bar{x}_{...})^2 \right] = 2N \text{Var}(\bar{x}_{k..}). \quad (6.1)$$

To find the expected mean square $E[MS_b]$ in terms of n_j , $\text{Var}(\bar{x}_{k..})$ is expanded such that

$$\text{Var}(\bar{x}_{k..}) = \text{Var} \left(\frac{\sum_{i=1}^N \sum_{j=1}^2 x_{kij}}{2N} \right) = \frac{1}{4N^2} \text{Var} \left(\sum_{i=1}^N \sum_{j=1}^2 x_{kij} \right). \quad (6.2)$$

Now,

$$\begin{aligned} \text{Var} \left(\sum_{i=1}^N \sum_{j=1}^2 x_{kij} \right) &= \text{Var} \left(\sum_{i=1}^N x_{ki1} + \sum_{i=1}^N x_{ki2} \right) \\ &= \text{Var} \left(\sum_{i=1}^N x_{ki1} \right) + \text{Var} \left(\sum_{i=1}^N x_{ki2} \right) + 2\text{Cov} \left(\sum_{i=1}^N x_{ki1}, \sum_{i=1}^N x_{ki2} \right), \end{aligned} \quad (6.3)$$

but as the variables x_{ki1} and x_{ki2} represent the paternally and maternally inherited genes, they are independent for this model due to the assumption of random mating.

The covariance term in Expression (6.3) is then equal to zero.

So from Expressions (6.1) to (6.3),

$$E[MS_b] = \frac{1}{2N} \left(\text{Var} \left(\sum_{i=1}^N x_{ki1} \right) + \text{Var} \left(\sum_{i=1}^N x_{ki2} \right) \right) \quad (6.4)$$

and to evaluate the expected mean square fully the variance of the scores of paternally and maternally inherited genes within rots is calculated. First consider the paternally inherited allele, $j = 1$,

$$\text{Var} \left(\sum_{i=1}^N x_{ki1} \right) = \sum_{i=1}^N \text{Var}(x_{ki1}) + \sum_{i=1}^N \sum_{\substack{i'=1 \\ i' \neq i}}^N \text{Cov}(x_{ki1}, x_{ki'1}), \quad (6.5)$$

where

$$\sum_{i=1}^N \text{Var}(x_{ki1}) = \sum_{i=1}^N E[x_{ki1} - q]^2 = N q(1 - q), \quad (6.6)$$

and q is the population frequency of allele A_1 , i.e. $\Pr(x_{kij} = 1) = q$. The covariance term is more complex and depends on the relationship between the individuals in the sample:

$$\begin{aligned} \text{Cov}(x_{ki1}, x_{ki'1}) &= E[x_{ki1} x_{ki'1}] - E[x_{ki1}]E[x_{ki'1}] \\ &= 1 \cdot \Pr(x_{ki'1} = 1 \mid x_{ki1} = 1) \Pr(x_{ki1} = 1) - \Pr(x_{ki1} = 1)^2 \\ &= q (\Pr(x_{ki'1} = 1 \mid x_{ki1} = 1) - q). \end{aligned} \quad (6.7)$$

Details are now given for determining $\Pr(x_{ki'1} = 1 \mid x_{ki1} = 1)$, i.e. the probability that an allele is type A_1 given that a second allele from an individual drawn from the same rot is also type A_1 , where both alleles are of paternal origin.

1. If $i = i'$, then $\Pr(x_{ki'1} = 1 \mid x_{ki1} = 1) = 1$, and $\text{Cov}(x_{ki1}, x_{ki'1}) = q(1 - q)$.
2. If $i \neq i'$, then for $j = 1$ (the paternally contributed gene) let there be M males in the parental population.

- (a) If i and i' are sibs then, by definition, their paternally inherited alleles are from the same male. The probability that one allele is A_1 given that the other is A_1 is equal to the probability that the alleles are identical by descent ($\frac{1}{2}$) plus the probability that they are identical in state, but not identical by descent ($\frac{q}{2}$). The covariance, $\text{Cov}(x_{ki1}, x_{ki'1})$, is therefore equal to

$$q \left(\frac{1}{2} + \frac{q}{2} - q \right) = \frac{1}{2} q(1 - q).$$

For an individual i , there are $(N/M - 1)$ individuals i' who are sibs.

- (b) If i and i' are not sibs, the probability that both the alleles are A_1 is equal to the probability that the alleles are identical in state (q), and thus

$\text{Cov}(x_{ki1}, x_{ki'1}) = 0$. For an individual i , there are $(N - N/M)$ individuals i' who are not sibs.

So looking again at the covariance term from Expression (6.5),

$$\begin{aligned} \sum_{i=1}^N \sum_{\substack{i'=1 \\ i' \neq i}}^N \text{Cov}(x_{ki1}, x_{ki'1}) &= N \left(\frac{N}{M} - 1 \right) \cdot \frac{1}{2} q(1 - q) + N \left(N - \frac{N}{M} \right) \cdot 0 \\ &= \frac{1}{2} N q(1 - q) \left(\frac{N}{M} - 1 \right). \end{aligned} \quad (6.8)$$

Returning to Expression (6.5), using Expressions 6.6 and 6.8 yields

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^N x_{ki1}\right) &= N q(1 - q) + \frac{1}{2} N q(1 - q) \left(\frac{N}{M} - 1 \right) \\ &= \frac{1}{2} N q(1 - q) \left(1 + \frac{N}{M} \right). \end{aligned} \quad (6.9)$$

Similarly, as the allele frequencies are equal in males and females, for this model for the maternally inherited gene $j = 2$, so if there are F female parents

$$\sum_{i=1}^N \sum_{\substack{i'=1 \\ i' \neq i}}^N \text{Cov}(x_{ki2}, x_{ki'2}) = \frac{1}{2} N q(1 - q) \left(\frac{N}{F} - 1 \right), \quad (6.10)$$

and

$$\text{Var}\left(\sum_{i=1}^N x_{ki2}\right) = \frac{1}{2} N q(1 - q) \left(1 + \frac{N}{F} \right).$$

The expected mean square between rots is now, from Expressions (6.4) and (6.9),

$$\begin{aligned} E[\text{MS}_b] &= \frac{1}{2N} \left[\frac{1}{2} N q(1 - q) \left(1 + \frac{N}{M} \right) + \frac{1}{2} N q(1 - q) \left(1 + \frac{N}{F} \right) \right] \\ &= \frac{1}{4} q(1 - q) \left(2 + \frac{N}{M} + \frac{N}{F} \right) \end{aligned} \quad (6.11)$$

where M and F are the number of males and females from the parental generation that contribute alleles to the offspring in the sample.

Alternatively, as $N/M = N/F = n_f$ in the simple model, the expected mean square between rots now becomes

$$E[\text{MS}_b] = \frac{1}{2} q(1 - q)(1 + n_f) \quad (6.12)$$

where q is the frequency of the allele A_1 for which the indicator variable takes value 1 and n_f is the number of sibs per family. The primary interest here is to find an estimate of n_f under the simple model, so Expression (6.12) is used in preference to Expression (6.11).

6.3.2 The Expected Mean Square Within Rots

The expected mean squared deviation from the mean within rots is given in Table 6.1 as

$$E[MS_w] = \frac{1}{R(2N-1)} E \left[\sum_{k=1}^R \sum_{i=1}^N \sum_{j=1}^2 (x_{kij} - \bar{x}_{k..})^2 \right]. \quad (6.13)$$

To evaluate this expression in terms of n_f , the first step is to find $E[(x_{kij} - \bar{x}_{k..})^2]$. Now,

$$\text{Var}(x_{kij} - \bar{x}_{k..}) = E[(x_{kij} - \bar{x}_{k..})^2] - (E[x_{kij} - \bar{x}_{k..}])^2$$

but $E[x_{kij} - \bar{x}_{k..}] = 0$, therefore

$$E[(x_{kij} - \bar{x}_{k..})^2] = \text{Var}(x_{kij} - \bar{x}_{k..}) = \text{Var}(x_{kij}) + \text{Var}(\bar{x}_{k..}) - 2\text{Cov}(x_{kij}, \bar{x}_{k..}). \quad (6.14)$$

It was shown previously (Section 6.3.1) that

$$\text{Var}(\bar{x}_{k..}) = \frac{1}{4N} q(1-q)(1+n_f)$$

and that $\text{Var}(x_{kij}) = q(1-q)$. The covariance term in Expression (6.14) expands to give

$$\text{Cov}(x_{kij}, \bar{x}_{k..}) = \frac{1}{2N} \text{Cov} \left(x_{kij}, \sum_{i=1}^N \sum_{j=1}^2 x_{kij} \right). \quad (6.15)$$

The covariance of x_{kij} and $x_{ki'j'}$ is therefore as follows:

1. if $j \neq j'$, for this model the value of $x_{ki'j'}$ is independent of the value of x_{kij} and

$$\Pr(x_{ki'j'} = 1 \mid x_{kij} = 1) = \Pr(x_{ki'j'} = 1) = q.$$

This is true whether the two alleles concerned are from the same or different individuals. The covariance term when $j \neq j'$ is $\text{Cov}(x_{kij}, x_{ki'j'}) = q \cdot q - q^2 = 0$ (see Expression (6.7)).

2. If $j = j'$ (both alleles are inherited from a parent of the same sex) then

$$(a) \text{ if } i = i', \text{Cov}(x_{kij}, x_{ki'j}) = q(1-q);$$

$$(b) \text{ if } i \neq i', \text{ and } i \text{ and } i' \text{ are siblings then there will be } (n_f - 1) \text{ pairs } (x_{kij}, x_{ki'j}) \text{ such that } \text{Cov}(x_{kij}, x_{ki'j}) = \frac{1}{2}q(1-q). \text{ If } i \text{ and } i' \text{ are not sibs, there will be } (N - n_f) \text{ pairs } (x_{kij}, x_{ki'j}) \text{ such that } \text{Cov}(x_{kij}, x_{ki'j}) = 0.$$

Substituting these covariances into Expression (6.15)

$$\text{Cov}(x_{kij}, \bar{x}_{k..}) = \frac{1}{2N} \left(q(1-q) + (n_f - 1) \frac{1}{2} q(1-q) \right) = \frac{1}{4N} q(1-q)(1+n_f),$$

and returning to Expression (6.14)

$$\begin{aligned} E[(x_{kij} - \bar{x}_{k..})^2] &= q(1-q) + \frac{1}{4N} q(1-q)(1+n_f) - \frac{1}{2N} q(1-q)(1+n_f) \\ &= q(1-q) \left(1 - \frac{(1+n_f)}{4N} \right). \end{aligned} \quad (6.16)$$

From this

$$E \left[\sum_{k=1}^R \sum_{i=1}^N \sum_{j=1}^2 (\bar{x}_{kij} - x_{k..})^2 \right] = 2NRq(1-q) \left(1 - \frac{(1+n_f)}{4N} \right),$$

and the expected mean square within rots is

$$\begin{aligned} E[\text{MS}_w] &= \frac{1}{R(2N-1)} 2NRq(1-q) \left(1 - \frac{(1+n_f)}{4N} \right) \\ &= q(1-q) \left(1 - \frac{(n_f-1)}{2(2N-1)} \right). \end{aligned} \quad (6.17)$$

The Between Rot Variance, σ_b^2

The expected mean squares are evaluated in terms of n_f in Sections 6.3.1 and 6.3.2 and found to be

$$E[\text{MS}_b] = \frac{1}{2} q(1-q)(1+n_f) \quad \text{and} \quad E[\text{MS}_w] = q(1-q) \left(1 - \frac{(n_f-1)}{2(2N-1)} \right)$$

(Expressions (6.12) and (6.17)). The between rot variance σ_b^2 can be found from these expressions; $E[\text{MS}_b] = \sigma_w^2 + 2N\sigma_b^2$ and $E[\text{MS}_w] = \sigma_w^2$. Thus

$$\sigma_b^2 = \frac{1}{2N} (E[\text{MS}_b] - E[\text{MS}_w]) = \frac{q(1-q)(n_f-1)}{2(2N-1)}. \quad (6.18)$$

(As a check, note that the total variance $\sigma_w^2 + \sigma_b^2 = q(1-q)$ as expected). If $n_f = 1$, the within rot variance equals $q(1-q)$ (the variance of X_{kij}) and the between rot variance is zero. As n_f increases, the between rot variance increases and the within rot variance decreases as individuals at each rot tend to be more like each other than flies from another rot.

The ratio $E[MS_b]/E[MS_w]$

Suppose an experiment is carried out in which N individuals are sampled from each of R rots; an analysis of variance is carried out based on the observed genotypes of the RN individuals, with $x_{kij} = 0, 1$ as described above. Let F_A be the observed ratio of the mean squares (MS_b/MS_w). This observed value is compared with the expected ratio of mean squares to obtain an estimate of n_f .

The ratio of the expected mean squares found from the analysis of variance is independent of q and R , namely

$$E[MS_b]/E[MS_w] = \frac{\frac{1}{2}q(1-q)(1+n_f)}{q(1-q)(1-((n_f-1)/2(2N-1)))} = \frac{(2N-1)(n_f+1)}{2(2N-1)-n_f+1}. \quad (6.19)$$

However, the ratio of the expected mean squares $E[MS_b]/E[MS_w]$ is not equal to the expected ratio $E[(MS_b)/(MS_w)]$. If normality is assumed for a preliminary investigation, the ratio $(MS_b/E[MS_b])/(MS_w/E[MS_w])$ is known to have an F distribution and thus the expected value is $df_w/(df_w-2)$, where df_w is the degrees of freedom for the denominator (MS_w). Hence,

$$E \left[\frac{MS_b/E[MS_b]}{MS_w/E[MS_w]} \right] = \frac{R(2N-1)}{R(2N-1)-2} \quad (6.20)$$

and

$$E[F_A] = E \left[\frac{MS_b}{MS_w} \right] = \left(\frac{R(2N-1)}{R(2N-1)-2} \right) \left(\frac{(2N-1)(n_f+1)}{2(2N-1)-n_f+1} \right) \quad (6.21)$$

The estimate for family size can be found by rearrangement. If N or R are large, then $R(2N-1)/(R(2N-1)-2)$ is approximately equal to one and $E[F_A] \approx E[MS_b]/E[MS_w]$. With this approximation,

$$\hat{n}_f = \frac{4NF_A}{F_A + 2N - 1} - 1. \quad (6.22)$$

The expected value of F_A is given in Expression (6.21). To find the variance of F_A , note that

$$\text{Var} \left[\frac{MS_b/E[MS_b]}{MS_w/E[MS_w]} \right] = \frac{2df_w^2(df_w + df_b - 2)}{df_b(df_w - 2)^2(df_w - 4)} \quad (6.23)$$

where $df_w = R(2N-1)$ and $df_b = R-1$. Hence,

$$\text{Var} \left[\frac{MS_b}{MS_w} \right] = \left(\frac{2(R(2N-1))^2(R(2N-1) + (R-1) - 2)}{(R-1)(R(2N-1)-2)^2(R(2N-1)-4)} \right) \left(\frac{E[MS_b]}{E[MS_w]} \right)^2 \quad (6.24)$$

If N or R are assumed to be large, then $df_w - 2 \approx df_w$, $df_w - 4 \approx df_w$ and $df_w + df_b - 2 \approx df_w + df_b$. Thus,

$$\text{Var} \left[\frac{MS_b/E[MS_b]}{MS_w/E[MS_w]} \right] \approx \frac{2}{df_w} + \frac{2}{df_b} = \frac{2}{R-1} + \frac{2}{R(2N-1)}$$

From this

$$\text{Var}(F_A) = \text{Var} \left[\frac{\text{MS}_b}{\text{MS}_w} \right] \approx \left(\frac{(2N-1)(n_f+1)}{2(2N-1)-n_f+1} \right)^2 \left(\frac{2}{R-1} + \frac{2}{R(2N-1)} \right). \quad (6.25)$$

For a given N and R , the variance of F_A increases with increasing n_f . If $n_f = 1$, then the variance of F_A is a function of N and R , whereas for $n_f > 1$ the variance of F_A is a function of n_f^2 as well.

These expressions for $\text{Var}(F_A)$ and $\hat{n}_f$ are compared with empirical values in the next section.

6.3.3 Results from Simulations Experiments

In the previous section, the expected mean squares from an analysis of variance of the indicator variable x_{kij} are found in terms of the family size, n_f . An estimate, $\hat{n}_f$, of the family size is found using results from an analysis of sampled genotype data. If genotype data are generated for individuals with known relatedness, then the value of $\hat{n}_f$ can be compared with the true n_f to assess the reliability of the method. In this section, the mean squares, MS_b and MS_w , are found for a number of replicate trials using simulated data and the values compared with the expected values shown in Expressions (6.12) and (6.17). The ratio of the mean squares, F_A , is used to calculate $\hat{n}_f$ from Expression (6.22) for a range of simulated family sizes.

For each replicate trial, appropriate genotypes are simulated for RN individuals according to the population structure defined by the model described in Section 6.2. An analysis of variance is carried out, and the observed mean squares within and between rots recorded. This is achieved using a computer program written specifically for this purpose (see Appendix B.2).

The Mean Squares MS_b and MS_w

Let there be 10 rots, with 10 individuals sampled from each rot ($N = R = 10$). Within this basic design, let the number of parent pairs (P) colonising each rot be either 1, 5 or 10; thus the true (simulated) n_f is 10, 2 or 1, respectively. Each trial is repeated 20 times for each n_f with a population gene frequency, q , of 0.1, 0.3 and 0.5. The

theoretical $E[MS_b]$ is calculated from Expression (6.12) and the theoretical $E[MS_w]$ is calculated from Expression (6.17) for the appropriate allele frequencies and family size.

Figure 6.2 shows MS_b and MS_w as calculated for each trial for $n_f = 10, 2$ and 1 . Each graph shows results for $q = 0.1, 0.3$ and 0.5 as indicated by the different symbols; the theoretical values for the given parameters are shown by the large symbol for each set of points. The line $MS_b = MS_w$ has been added to aid comparison between figures. When the simulated family size is equal to one, the individuals sampled from the rots are essentially unrelated and thus represent a random sample of genotypes. The variance between rots, σ_b^2 , becomes zero and the ratio of mean squares is approximately one. When the simulated family size is ten, for this example each rot hosts one family and as expected the variance between rots is much larger than the variance within rots. The variance of observed values from different trials is large when n_f is large.

The Family Size Estimator, $\hat{n}_f$

Using the results from the analysis of variance of simulated data described above, an estimate of family size is found for each replicate trial. A total of 500 trials are carried out for each set of parameters. The mean $\hat{n}_f$ taken over all trials is then compared with the true value. From Expression (6.22),

$$\hat{n}_f \approx \frac{4NF_A}{F_A + 2N - 1} - 1 \quad \text{for large } N \text{ or } R$$

where N is the total number of offspring sampled per rot, R the number of rots and F_A is the observed ratio MS_b/MS_w . The variance of $\hat{n}_f$ can be found using the intraclass correlation, $t = \sigma_b^2/\sigma_b^2 + \sigma_w^2$. From Expressions (6.17) and (6.18),

$$t = \frac{n_f - 1}{2(2N - 1)} \quad (6.26)$$

and hence the estimator of n_f can be written in terms of t ,

$$\hat{n}_f = 2(2N - 1)t + 1. \quad (6.27)$$

The variance of t is known [Fisher, 1941],

$$\text{Var}(t) = \frac{2(1-t)^2(1+(2N-1)t)^2}{(R-1)2N(2N-1)}. \quad (6.28)$$

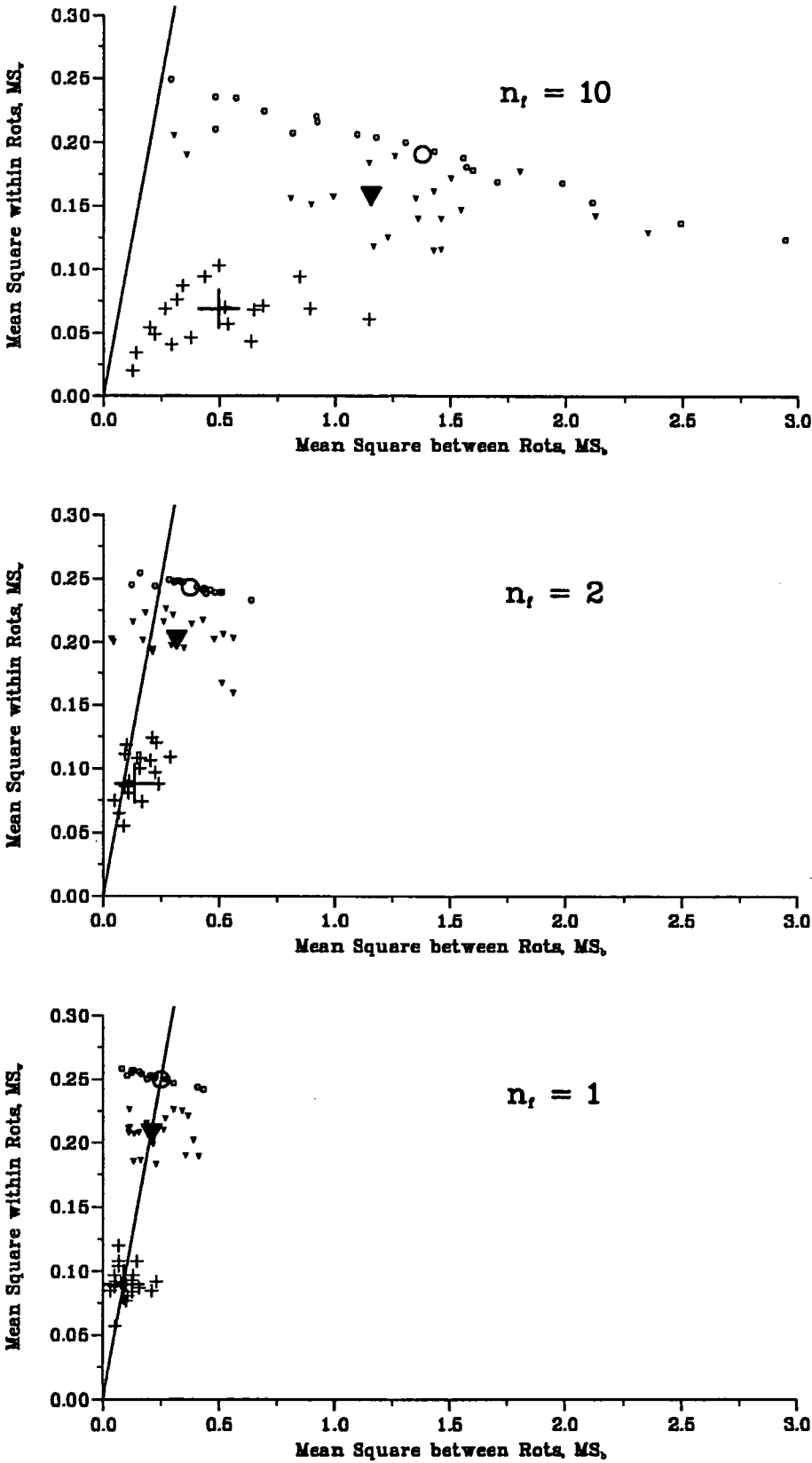


Figure 6.2: The observed MS_b and MS_w from an analysis of variance based on allele scores, for $n_f = 10, 2$ or 1 and $q = 0.5(\circ), 0.3(\nabla)$ and $0.1(+)$. The large symbol in each graph indicates the expected values and the line represents $MS_b = MS_w$.

Simulated n_f	Expected $\text{Var}(\hat{n}_f)$	$q = 0.1$		$q = 0.5$	
		mean $\hat{n}_f$	$\text{Var}(\hat{n}_f)$	mean $\hat{n}_f$	$\text{Var}(\hat{n}_f)$
1	0.845	0.993	0.702	0.990	0.826
2	1.801	1.959	1.476	2.080	1.809
5	6.086	4.943	6.330	5.082	6.523
10	14.876	9.596	14.738	9.772	16.330

Table 6.2: Results obtained from 500 replicate simulation experiments for each set of parameters, for data based on allele scores. Expected results are calculated for $N = R = 10$ and for the n_f and allele frequency as shown.

From Expression 6.28, the variance of the estimator n_f is

$$\begin{aligned}\text{Var}(\hat{n}_f) &= 4(2N - 1)^2 \text{Var}(t) \\ &= \frac{8(2N - 1)(1 - t)^2(1 + (2N - 1)t)^2}{(R - 1)2N}.\end{aligned}\quad (6.29)$$

Substituting t from Expression (6.26),

$$\begin{aligned}\text{Var}(\hat{n}_f) &= \frac{4(2N - 1)(4N - n_f - 1)^2(1 + (n_f - 1)/2)^2}{N(R - 1)(2(2N - 1))^2} \\ &= \frac{(4N - n_f - 1)^2(n_f + 1)^2}{4N(R - 1)(2N - 1)}.\end{aligned}\quad (6.30)$$

Table 6.2 shows both the expected results and those obtained from simulations for the mean $\hat{n}_f$ and $\text{Var}(\hat{n}_f)$; the true family size is 1, 2, 5 or 10. The mean estimated family size lies close to the true value for any gene frequency, although only results for $q = 0.5$ and $q = 0.1$ are shown. These simulation results are also shown in Figures 6.3 and 6.4, respectively. The estimated $\hat{n}_f$ can be less than 0 or greater than N ; all values less than 0.5 were classed as family size zero by the analysis program. The observed variance of $\hat{n}_f$ is close to that expected from Expression (6.30) but note that the variance is very high when the family size is large. Consider the magnitude of this variance as N and R increase.

- If $n_f = 1$, the sample essentially consists of N unrelated individuals and

$$\begin{aligned}\text{Var}(\hat{n}_f) &= \frac{(4N - 2)^2(2)^2}{4N(R - 1)(2N - 1)} = \frac{1}{R - 1}\left(8 - \frac{4}{N}\right) \\ &\approx \frac{6}{(R - 1)} \text{ for } N = 2, \text{ and} \\ &\approx \frac{8}{(R - 1)} \text{ for large } N.\end{aligned}$$

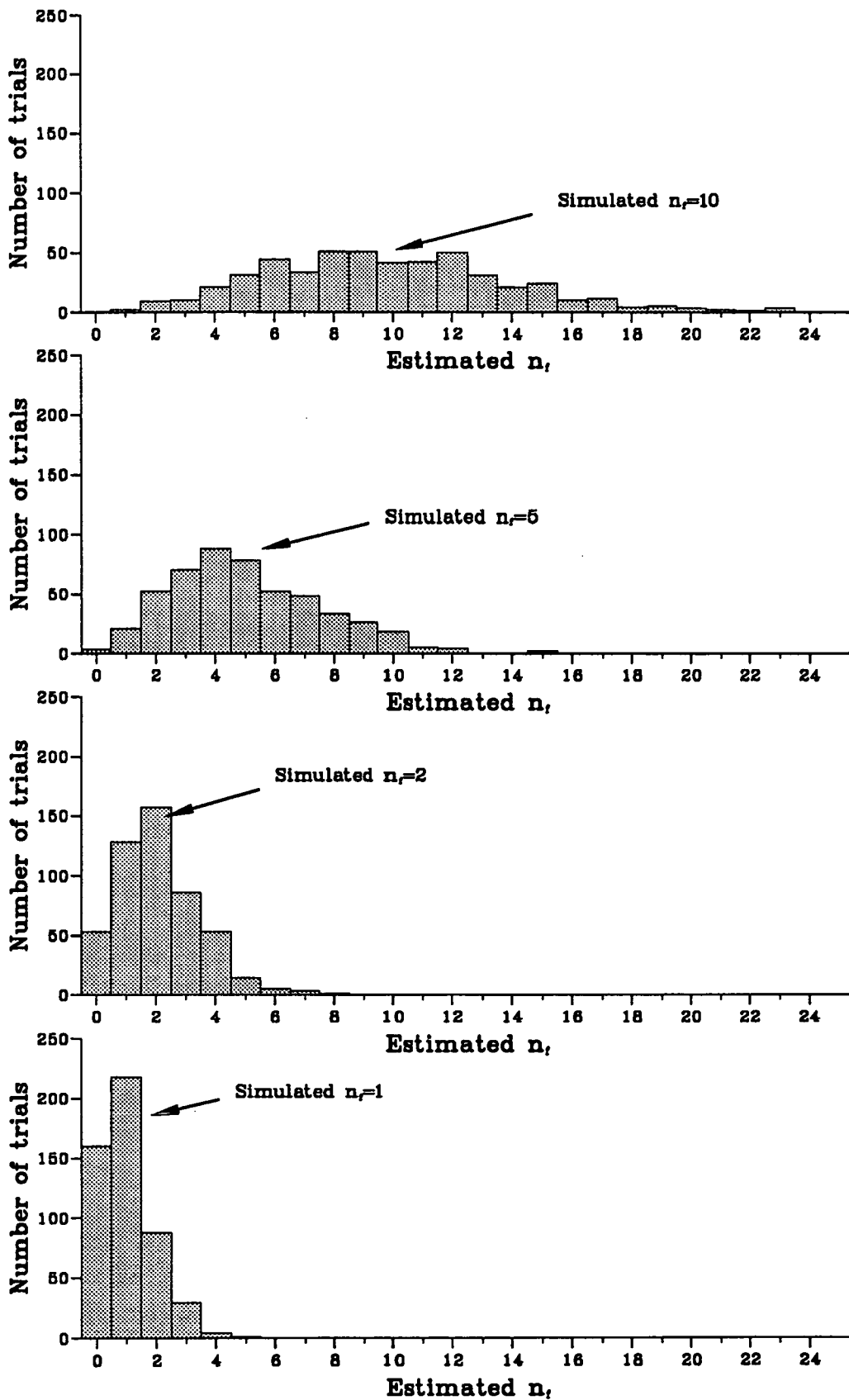


Figure 6.3: The estimates of family size, $\hat{n}_f$, for simulated $n_f = 1, 2, 5$ and 10, with $q = 0.5$. There are 500 replicates trials with $N = R = 10$ for each n_f , using data based on allele scores.

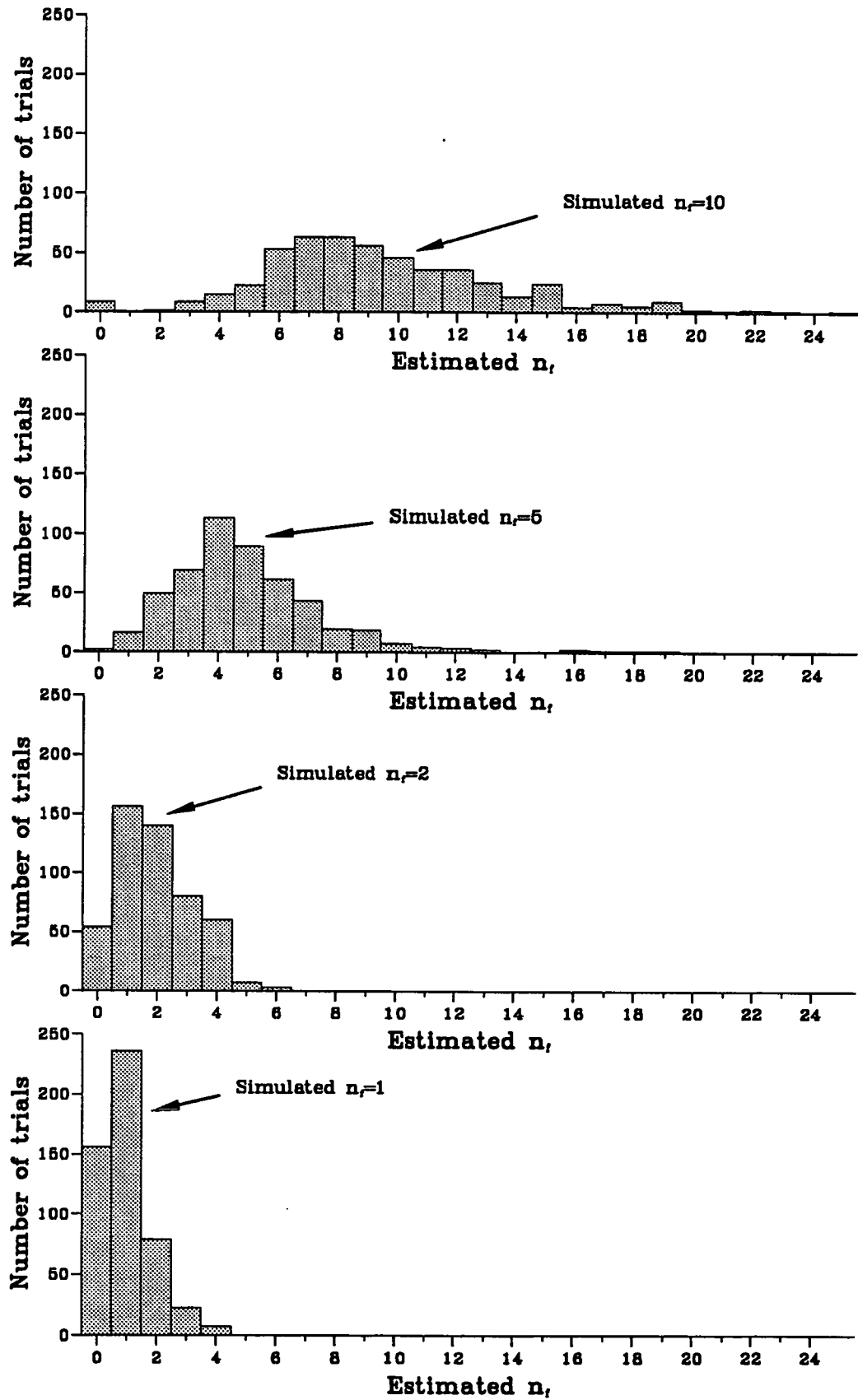


Figure 6.4: The estimates of family size, $\hat{n}_f$, for simulated $n_f = 1, 2, 5$ and 10 , with $q = 0.1$. There are 500 replicates trials with $N = R = 10$ for each n_f , using data based on allele scores.

- If $n_f = N$, the sample for each rot consists of one family of N sibs,

$$\begin{aligned}\text{Var}(\hat{n}_f) &= \frac{(3N-1)^2(N+1)}{4N(R-1)(2N-1)} \\ &\approx \frac{25}{8(R-1)} \text{ for } N=2, \text{ and} \\ &\approx \frac{9N^2}{8(R-1)} \text{ for large } N.\end{aligned}$$

Results from repeated simulations agree with those expected, especially when the family size is small.

6.4 Analysis of Variance Based on Genotype Scores

In this section, the analysis of variance is carried out using an indicator variable based on genotype rather than allelic data. Estimates of family size are found in a similar way to that described in Section 6.3, and values for the mean and variance of $\hat{n}_f$ are compared with those expected.

Let each individual in the sample have a known genotype for a single locus that has two alleles. Each rot now has a total of N scores associated with it. Let

$$x_{ij} = \begin{cases} 0 & \text{if genotype is } A_1A_1 \\ 1 & \text{if genotype is } A_1A_2 \\ 2 & \text{if genotype is } A_2A_2, \end{cases}$$

where $i = 1, 2, \dots, R$, the number of rots, and $j = 1, 2, \dots, N$, the number of individuals per rot.

6.4.1 The Expected Mean Square Between Rots

From Table 6.3, the expected mean squared deviation from the mean between rots is

$$E[\text{MS}_b] = \frac{N}{R-1} E \left[\sum_{i=1}^R (\bar{x}_i - \bar{x}_{..})^2 \right] = \frac{N}{R-1} [(R-1)\text{Var}(\bar{x}_i)]. \quad (6.31)$$

The expansion of $\text{Var}(\bar{x}_i)$ is as follows;

$$\text{Var}(\bar{x}_i) = \frac{1}{N^2} \text{Var} \left(\sum_{j=1}^N x_{ij} \right) = \frac{1}{N^2} \left(\sum_{j=1}^N \text{Var}(x_{ij}) + \sum_{j=1}^N \sum_{\substack{j'=1 \\ j' \neq j}}^N \text{Cov}(x_{ij}, x_{ij'}) \right). \quad (6.32)$$

Source of Variation	Degrees of Freedom	Sum of Squares
between rots	$R - 1$	$N \sum_{i=1}^R (\bar{x}_{i.} - \bar{x}_{..})^2$
within rots	$R(N - 1)$	$\sum_{i=1}^R \sum_{j=1}^N (x_{ij} - \bar{x}_{i.})^2$
Total	$RN - 1$	$\sum_{i=1}^R \sum_{j=1}^N (x_{ij} - \bar{x}_{..})^2$

Table 6.3: The analysis of variance table using data based on genotype scores.

For this analysis the expected value of x_{ij} is

$$E[x_{ij}] = \sum_{j=1}^N x_{ij} \Pr(x_{ij}) = 0q^2 + 1(2q(1 - q)) + 2(1 - q)^2 = 2(1 - q) \quad (6.33)$$

where q is the population frequency of allele A_1 . The variance of each observation is therefore

$$\begin{aligned} \text{Var}(x_{ij}) &= E[x_{ij} - E[x_{ij}]]^2 \\ &= 2q(1 - q) + 4(1 - q)^2 + 4(1 - q)^2 - 4(1 - q)(2(1 - q)) \\ &= 2q(1 - q). \end{aligned} \quad (6.34)$$

The covariance term in Expression (6.32) depends on the relationship between the individuals in the sample;

1. if $j \neq j'$ and j' is not a sib of j then

$$\begin{aligned} E[x_{ij}x_{ij'}] &= \sum (x_{ij}x_{ij'} \Pr(x_{ij}) \Pr(x_{ij'})) \\ &= 1(2q(1 - q) 2q(1 - q)) + 4(2q(1 - q)(1 - q)^2) + 4((1 - q)^2(1 - q)^2) \\ &= 4(1 - q)^2, \end{aligned} \quad (6.35)$$

$$\text{and } \text{Cov}(x_{ij}, x_{ij'}) = E[x_{ij}x_{ij'}] - E[x_{ij}]E[x_{ij'}] = 4(1 - q)^2 - (2(1 - q))^2 = 0.$$

2. If $j \neq j'$ and j' is a full sib of j then

$$E[x_{ij}x_{ij'}] = \sum (x_{ij}x_{ij'} \Pr(x_{ij}) \Pr(x_{ij'}))$$

$$\begin{aligned}
&= 1(q(1-q)(1+q(1-q))) + 4\left(\frac{1}{2}q(1-q)^2(2-q)\right) \\
&\quad + 4\left(\frac{1}{4}(1-q)^2(2-q)^2\right) \\
&= (1-q)(1+3(1-q)), \tag{6.36}
\end{aligned}$$

$$\text{and } \text{Cov}(x_{ij}, x_{ij'}) = (1-q)(4-3q) - (2(1-q))^2 = q(1-q).$$

3. If $j = j'$ then

$$\text{Cov}(x_{ij}, x_{ij'}) = E[x_{ij}^2] - (E[x_{ij}])^2 = 2q(1-q). \tag{6.37}$$

So returning to Expression (6.32),

$$\begin{aligned}
\text{Var}(\bar{x}_{i.}) &= \frac{1}{N^2} (N2q(1-q) + N(N-n_f)(0) + N(n_f-1)(q(1-q))) \\
&= \frac{1}{N} q(1-q)(1+n_f). \tag{6.38}
\end{aligned}$$

The expected mean square between rots, Expression (6.31), is now

$$E[\text{MS}_b] = q(1-q)(1+n_f), \tag{6.39}$$

where q is the frequency of the allele A_1 , and n_f is the family size. This is twice the value of $E[\text{MS}_b]$ for the analysis based on allele scores.

6.4.2 The Expected Mean Square Within Rots

From Table 6.3, the expected mean squared deviation from the mean within rots is

$$E[\text{MS}_w] = \frac{1}{R(N-1)} E \left[\sum_{i=1}^R \sum_{j=1}^N (x_{ij} - \bar{x}_{i.})^2 \right]. \tag{6.40}$$

Following a similar procedure to that described in Section 6.3.2, the expected mean square within rots is found in terms of n_f ;

$$E[(x_{ij} - \bar{x}_{i.})^2] = \text{Var}(x_{ij} - \bar{x}_{i.}) = \text{Var}(x_{ij}) + \text{Var}(\bar{x}_{i.}) - 2\text{Cov}(x_{ij}, \bar{x}_{i.}). \tag{6.41}$$

The covariance term expands to give,

$$\begin{aligned}
\text{Cov}(x_{ij}, \bar{x}_{i.}) &= \frac{1}{N} \text{Cov} \left(x_{ij}, \sum_{j=1}^N x_{ij} \right) \\
&= \frac{1}{N} (2q(1-q) + (N-n_f)(0) + (n_f-1)q(1-q)) \\
&= \frac{1}{N} q(1-q)(1+n_f), \tag{6.42}
\end{aligned}$$

and substituting these values into Expression (6.41) gives

$$\begin{aligned}
 E[x_{ij} - \bar{x}_{i.}]^2 &= \text{Var}(x_{ij} + \text{Var}(\bar{x}_{i.}) - 2\text{Cov}(x_{ij}, \bar{x}_{i.})) \\
 &= 2q(1 - q) + \frac{1}{N}q(1 - q)(1 + n_f) - \frac{2}{N}q(1 - q)(1 + n_f) \\
 &= 2q(1 - q) \left(1 - \frac{(1 + n_f)}{2N}\right). \tag{6.43}
 \end{aligned}$$

The expected mean square within rots is then

$$\begin{aligned}
 E[\text{MS}_w] &= \frac{1}{R(N - 1)} E \left[\sum_{i=1}^R \sum_{j=1}^N (x_{ij} - \bar{x}_{i.})^2 \right] = \frac{2NR}{R(N - 1)} q(1 - q) \left(1 - \frac{(1 + n_f)}{2N}\right) \\
 &= 2q(1 - q) \left(1 - \frac{(n_f - 1)}{2(N - 1)}\right). \tag{6.44}
 \end{aligned}$$

The Between Rot Variance, σ_b^2

The between rot variance, σ_b^2 , is found by combining expressions for the expected mean squares. From Expressions (6.39) and (6.44),

$$\sigma_b^2 = \frac{1}{N} (E[\text{MS}_b] - E[\text{MS}_w]) = 2q(1 - q) \frac{(n_f - 1)}{2(N - 1)}. \tag{6.45}$$

As expected, $\sigma_w^2 + \sigma_b^2 = 2q(1 - q)$. As before, the between rot variance is zero when $n_f = 1$ and σ_w^2 is $2q(1 - q)$. The σ_b^2 increases as n_f increases but σ_w^2 decreases.

The Ratio $E[\text{MS}_b]/E[\text{MS}_w]$

The ratio of the mean squares found from the analysis of variance is independent of q and R , and is very similar to the ratio found from the analysis of alleles scores (see Expression (6.19)), namely

$$E[\text{MS}_b]/E[\text{MS}_w] = \frac{(N - 1)(n_f + 1)}{2(N - 1) - n_f + 1}. \tag{6.46}$$

If F_G is the observed ratio of the mean squares, then the expected value of F_G is

$$E[F_G] = E \left[\frac{\text{MS}_b}{\text{MS}_w} \right] \approx \frac{(N - 1)(n_f + 1)}{2N - n_f - 1} \left(1 + \frac{2}{R(N - 1)}\right), \tag{6.47}$$

and the variance of F_G is

$$\text{Var} \left[\frac{\text{MS}_b}{\text{MS}_w} \right] \approx \left(\frac{(N - 1)(n_f + 1)}{2N - n_f - 1} \right)^2 \left(\frac{2}{R - 1} + \frac{2}{R(N - 1)} \right). \tag{6.48}$$

Using these new expressions, the family size is estimated from simulation experiments and the observed and expected results are compared.

		Expected			Observed	
		$\frac{E[MS_b]}{E[MS_w]}$	$E\left[\frac{MS_b}{MS_w}\right]$	$Var\left(\frac{MS_b}{MS_w}\right)$	mean F_G	$Var(F_G)$
(a)	$n_f = 1$	1.000	1.022	0.244	1.012	0.263
	2	1.588	1.623	0.616	1.635	0.721
	5	3.857	3.942	3.637	4.062	4.900
	10	11.000	11.244	29.572	12.335	69.806
(b)	5	3.068	3.068	2.111	3.012	2.198
	50	33.886	33.954	257.487	35.583	347.823

Table 6.4: The expected ratio of mean squares and results obtained from simulation experiments and for a range of parameters, from analysis of variance of genotype scores. For the example (a), there are 500 replicate samples each with $N = R = 10$. For (b) there are 1000 replicate samples with $N = 100, R = 10$.

6.4.3 Results from Simulation Experiments

The simulation experiments are carried out by generating genotypes for a number of individuals for a given population structure, according to the model described in Section 6.2. The value of the indicator variable x_{ij} is found for each individual and an analysis of variance carried out. Table 6.4 shows typical results obtained for the ratio MS_b/MS_w from these simulation experiments; the values shown are for $q = 0.5$ but results are similar for other gene frequencies. The ratio of the mean squares within and between rots obtained from simulation experiments is given by F_G ; Table 6.4 shows the expected ratio and the approximations (from Expressions (6.46), (6.47) and (6.48)). For the examples used here, the expected ratio of mean squares is close to the ratio of the expected mean squares and so the approximate expression shown in Expression (6.46) is used to find the estimate of n_f . The observed F_G is close to the expected values, but the variance of F_G is higher than expected particularly for large n_f .

Figure 6.5 shows MS_b and MS_w as calculated for 20 typical trials from example (a) in Table 6.4 for $n_f = 10, 2$ and 1. Each graph shows results for $q = 0.1, 0.3$ and 0.5 as indicated by the different symbols, and the theoretical value for the given parameters is shown by the large symbol for each set of points. The line $MS_b = MS_w$ has been added to aid comparison between figures. The theoretical $E[MS_b]$ is calculated

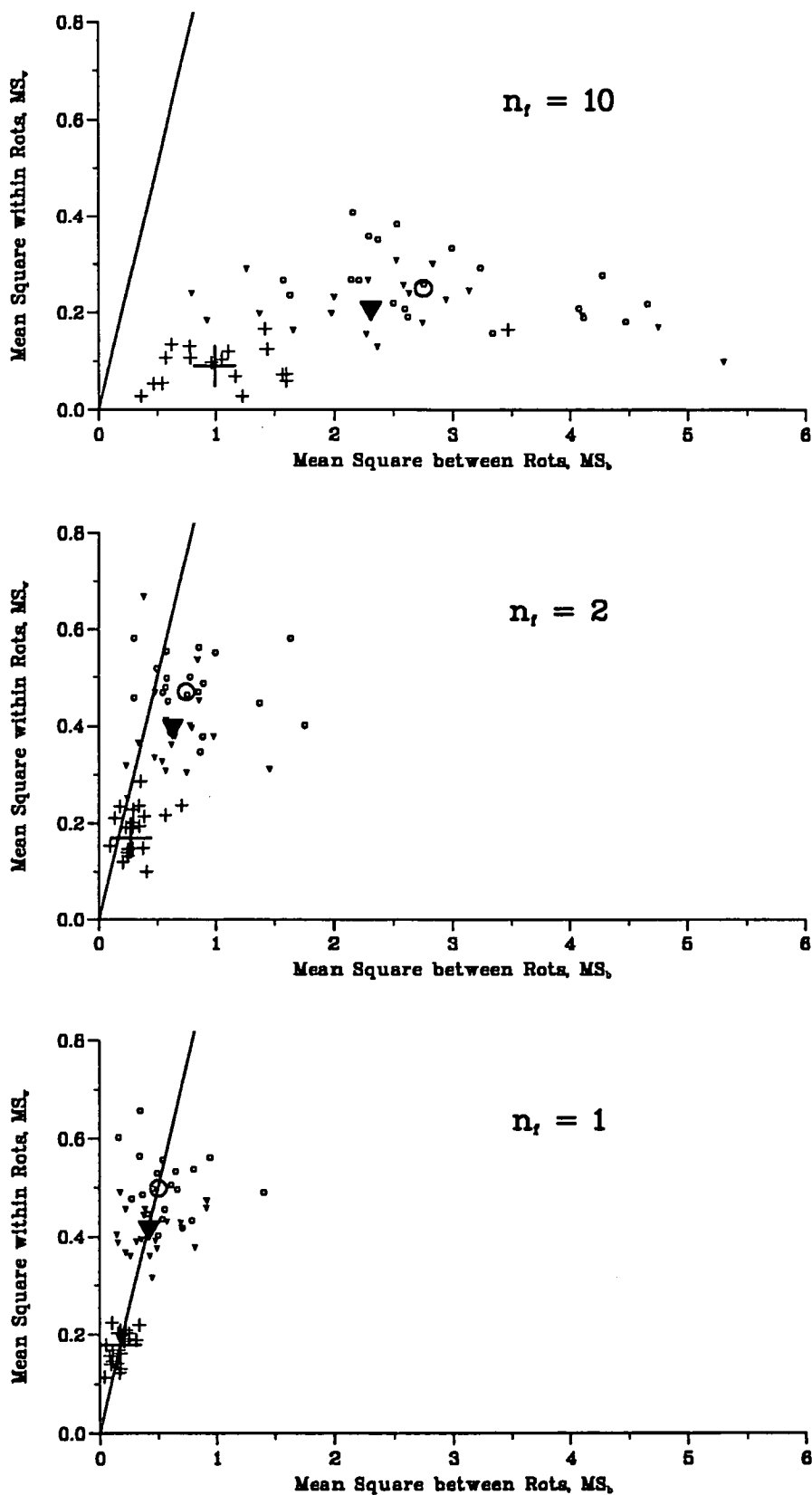


Figure 6.5: The observed MS_b and MS_w from an analysis of variance based on genotype scores, for $n_f = 10, 2$ or 1 and $q = 0.5(\circ), 0.3(\nabla)$ and $0.1(+)$. The large symbol in each graph indicates the expected values and the line represents $MS_b = MS_w$.

from Expression (6.12), and the theoretical $E[MS_w]$ is found from Expression (6.17) for the required allele frequencies and family sizes. These results are similar to those shown in Figure 6.2, although the experimental results are closer to the theoretical values when genotype data are used. As before, the expected ratio of MS_b and MS_w is independent of q so the large symbols (representing theoretical values) in Figure 6.5 lie on a straight line. The observed values cluster around the expected values but with increasing variance with increasing n_f .

The Estimator $\hat{n}_f$

From Expression (6.47), the estimate of family size is

$$\hat{n}_f \approx \frac{2F_G N}{N + F_G - 1} - 1. \quad (6.49)$$

The variance of the estimator $\hat{n}_f$ is found using the intraclass correlation as in Section 6.3, with

$$t = \sigma_b^2 / \sigma_b^2 + \sigma_w^2 = \frac{(n_f - 1)}{2(N - 1)},$$

namely

$$\text{Var}(\hat{n}_f) = \frac{(2N - n_f - 1)^2 (n_f + 1)^2}{2N(R - 1)(N - 1)}. \quad (6.50)$$

- If $n_f = 1$, the sample essentially consists of N unrelated individuals and

$$\begin{aligned} \text{Var}(\hat{n}_f) &= \frac{(2N - 2)^2 (2)^2}{2N(R - 1)(N - 1)} = \frac{1}{R - 1} \left(8 - \frac{8}{N}\right) \\ &\approx \frac{4}{(R - 1)} \text{ for } N = 2, \text{ and} \\ &\approx \frac{8}{(R - 1)} \text{ for large } N. \end{aligned}$$

- If $n_f = N$, the sample from each rot consists of 1 family of N sibs,

$$\begin{aligned} \text{Var}(\hat{n}_f) &= \frac{(N - 1)^2 (N + 1)^2}{2N(R - 1)(N - 1)} \\ &\approx \frac{9}{4(R - 1)} \text{ for } N = 2, \text{ and} \\ &\approx \frac{N^2}{2(R - 1)} \text{ for large } N. \end{aligned}$$

As for the analysis based on allele scores, the variance of $\hat{n}_f$ is proportional to N^2 when the family size is large. The estimated family size for each trial is found using

<i>N</i>	<i>R</i>	mean		Expected	Observed
		<i>n_f</i>	$\hat{n}_f$	Var($\hat{n}_f$)	Var($\hat{n}_f$)
100	10	100	96.996	560.648	775.678
100	10	50	49.724	324.036	356.492
100	10	25	24.873	114.854	123.743
100	10	20	19.733	79.299	83.393
100	10	10	9.797	24.256	23.678
100	10	5	4.864	7.601	7.695
100	10	4	4.052	5.336	5.267
100	10	2	2.045	1.960	2.056
100	10	1	1.044	0.880	0.904

Table 6.5: Results obtained from 500 replicate simulation experiments for each set of parameters, for data based on genotype scores.

Expression (6.49); the mean and variance of $\hat{n}_f$ obtained from simulation experiments are shown in Table 6.5 together with expected values. The mean $\hat{n}_f$ is very close to the true value for each example. The variance of $\hat{n}_f$ is close to the expected values when $\hat{n}_f$ is small but is greater than expected for large $\hat{n}_f$. Further examples are shown in Table 6.6, using different values of R and N . Here, results show that the variance of $\hat{n}_f$ is close to the expected value if n_f is small relative to N or R . Some typical results are also given in Figure 6.6; the parameters used in the simulations illustrated in this figure were the same as for those shown in Figure 6.3, enabling a direct comparison to be made between the two methods.

6.5 Analysis of Variance Based on Number of Heterozygotes

This section is a repeat of Sections 6.3 and 6.4, but with the analysis based on the number of heterozygotes. This analysis is carried out to investigate how the estimate of family size changes with different input data, and so enable identification of the most suitable form of data to use for further study. The expressions for the mean squares are obtained in a similar manner to that shown in Sections 6.3 and 6.4, so only the main results are stated in this section.

<i>N</i>	<i>R</i>	mean		Expected	Observed
		<i>n_f</i>	$\hat{n}_f$	Var($\hat{n}_f$)	Var($\hat{n}_f$)
10	10	1	1.003	0.799	0.741
10	20	1	0.995	0.379	0.346
10	50	1	1.006	0.147	0.137
10	100	1	1.008	0.073	0.069
10	10	5	4.996	4.356	4.360
10	20	5	5.033	2.062	2.062
10	50	5	4.979	0.799	0.922
10	100	5	4.971	0.396	0.446
10	10	10	9.606	6.052	7.829
10	20	10	9.781	2.866	3.356
10	50	10	9.885	1.111	1.302
10	100	10	9.950	0.551	0.661
5	10	1	0.979	0.711	0.717
10	10	1	1.003	0.799	0.741
20	10	1	0.979	0.845	0.780
50	10	1	1.006	0.870	0.852
100	10	1	1.044	0.880	0.904
10	10	10	9.606	6.052	7.829
100	10	10	9.797	24.256	23.678
200	10	10	9.783	25.563	26.708
500	10	10	9.955	26.358	28.356
1000	10	10	9.879	26.626	27.165
20	10	20	19.489	23.271	30.869
200	10	20	19.561	88.416	88.435
400	10	20	20.023	93.161	90.745
1000	10	20	19.790	96.040	93.258
2000	10	20	19.678	97.023	94.012

Table 6.6: Results obtained from simulation experiments for data based on genotype scores; for each set of parameters there are 500 replicate trials each with $q = 0.5$. For each trial, the genotypes for the individuals are generated for family size n_f ; R rots are used with N individuals analysed from each rot.

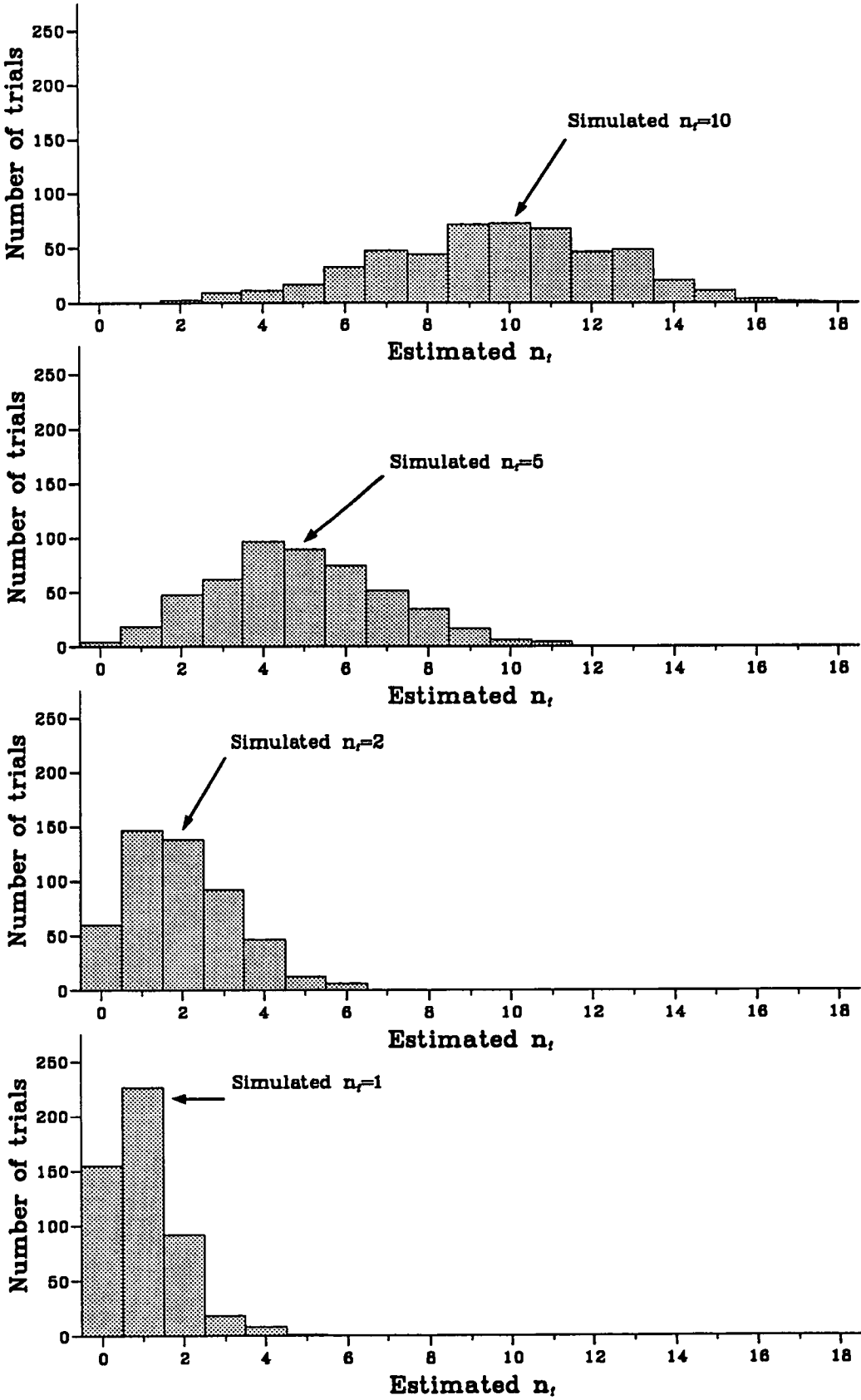


Figure 6.6: The estimates of family size, $\hat{n}_f$, for simulated $n_f = 1, 2, 5$ and 10 , with $q = 0.5$. There are 500 replicates trials with $N = R = 10$ for each n_f , using data based on genotype scores.

Source of Variation	Degrees of Freedom	Sum of Squares
between rots	$R - 1$	$N \sum_{i=1}^R (\bar{x}_{i.} - \bar{x}_{..})^2$
within rots	$R(N - 1)$	$\sum_{i=1}^R \sum_{j=1}^N (x_{ij} - \bar{x}_{i.})^2$
Total	$RN - 1$	$\sum_{i=1}^R \sum_{j=1}^N (x_{ij} - \bar{x}_{..})^2$

Table 6.7: The analysis of variance table using data based on number of heterozygotes.

Let each individual in the sample have a known genotype for a single locus that has two alleles. Each rot has a total of N scores associated with it. Let

$$x_{ij} = \begin{cases} 1 & \text{if genotype is } A_1A_2 \\ 0 & \text{otherwise,} \end{cases}$$

where $i = 1, 2, \dots, R$, the number of rots, and $j = 1, 2, \dots, N$, the number of offspring per rot.

6.5.1 The Expected Mean Square Between Rots

The expected value of x_{ij} is $2q(1 - q)$; let $2q(1 - q) = H$. The variance of x_{ij} is then $H(1 - H)$ and

$$\text{Var}(\bar{x}_{i.}) = \frac{H(1 - H)}{N} \left(1 + \frac{n_f - 1}{2} \left(1 - \frac{H}{2(1 - H)} \right) \right). \quad (6.51)$$

The expected mean square between rots (see Table 6.7) is

$$E[\text{MS}_b] = \frac{N}{R - 1} E \left[\sum_{i=1}^R (\bar{x}_{i.} - \bar{x}_{..})^2 \right] = q(1 - q)(1 + n_f - q(1 - q)(1 + 3n_f)) \quad (6.52)$$

where q is the frequency of the allele A_1 , and n_f is the family size.

6.5.2 The Expected Mean Square Within Rots

From Table 6.7, the expected mean squared deviation from the mean within rots is

$$E[MS_w] = \frac{1}{R(N-1)} E \left[\sum_{i=1}^R \sum_{j=1}^N (x_{ij} - \bar{x}_i)^2 \right]. \quad (6.53)$$

Using the indicator variable based on the number of heterozygotes, the expected mean square within rots becomes

$$\begin{aligned} E[MS_w] &= H(1-H) \left(1 - \frac{n_f - 1}{2} \left(1 - \frac{H}{2(1-H)} \right) \right) \\ &= q(1-q) \left(2(1-2q(1-q)) - \frac{n_f - 1}{N-1} (1-3q(1-q)) \right). \end{aligned} \quad (6.54)$$

The Between Rot Variance, σ_b^2

From Expressions (6.52) and (6.54), σ_b^2 is found to be

$$\sigma_b^2 = \frac{1}{N-1} q(1-q)(1-3q(1-q))(n_f - 1). \quad (6.55)$$

This is equal to zero for $n_f = 1$ as expected.

The Ratio $E[MS_b]/E[MS_w]$

In contrast to the previous examples, the ratio of the expected mean squares depends on the allele frequency;

$$\frac{E[MS_b]}{E[MS_w]} = \frac{(N-1)(1+n_f) - q(1-q)(N-1)(1+3n_f)}{(2N-n_f-1) + q(1-q)(3n_f-4N+1)}, \quad (6.56)$$

but is equal to 1 when $n_f = 1$ as before.

6.5.3 Results from Simulation Experiments

Sample genotypes are generated and the simulation experiments repeated as for the other datasets, with the programs modified to score heterozygote data. The value of the indicator variable x_{ij} is found for each individual and an analysis of variance carried out.

Figure 6.7 shows MS_b and MS_w as calculated for 20 trials, together with expected values found from Expressions (6.52) and (6.54). Unlike the previous examples, the ratio of MS_b and MS_w depends on q , so the large symbols (representing theoretical values) in Figure 6.7 do not lie on a straight line except when $n_f = 1$.

The Estimator n_f

Let F_H be the observed ratio MS_b/MS_w from analysis of simulated data. From Expression (6.56),

$$E[F_H] \approx \frac{(N-1)(1+n_f) - q(1-q)(N-1)(1+3n_f)}{(2N-n_f-1) - q(1-q)(4N-3n_f-1)}. \quad (6.57)$$

Using the intraclass correlation

$$t = \left(\frac{n_f - 1}{2(N-1)} \right) \left(\frac{2-3H}{2-2H} \right), \quad (6.58)$$

the estimate of family size, $\hat{n}_f$, can be written as

$$\hat{n}_f = 1 + \frac{4t(N-1)(1-H)}{2-3H}, \quad (6.59)$$

and the variance of $\hat{n}_f$ is

$$\text{Var}(\hat{n}_f) = \frac{16(N-1)^2(1-H)^2}{(2-3H)^2} \frac{2(1-t)^2(1+(N-1)t)^2}{N(N-1)(R-1)}. \quad (6.60)$$

Note that when $q = 0.5$, $H = 0.5$ and $\hat{n}_f$ is not dependent on allele frequency.

The observed ratio of mean squares, F_H , is used to find an estimate of n_f (from Expression (6.57)); typical results are shown in Figure 6.8.

6.6 Combining Results

When applied to the same data, the values obtained for $\hat{n}_f$ in Sections 6.3, 6.4 and 6.5 provide estimates for the family size. Clearly, these estimates are not independent. As an example, suppose the analyses based on genotypes and heterozygotes are applied to the same data; the estimates can then be compared directly.

Let there be 10 rots and 10 individuals sampled from each rot, with data from one locus with two alleles. Figure 6.9a shows the results obtained from analysis of 100

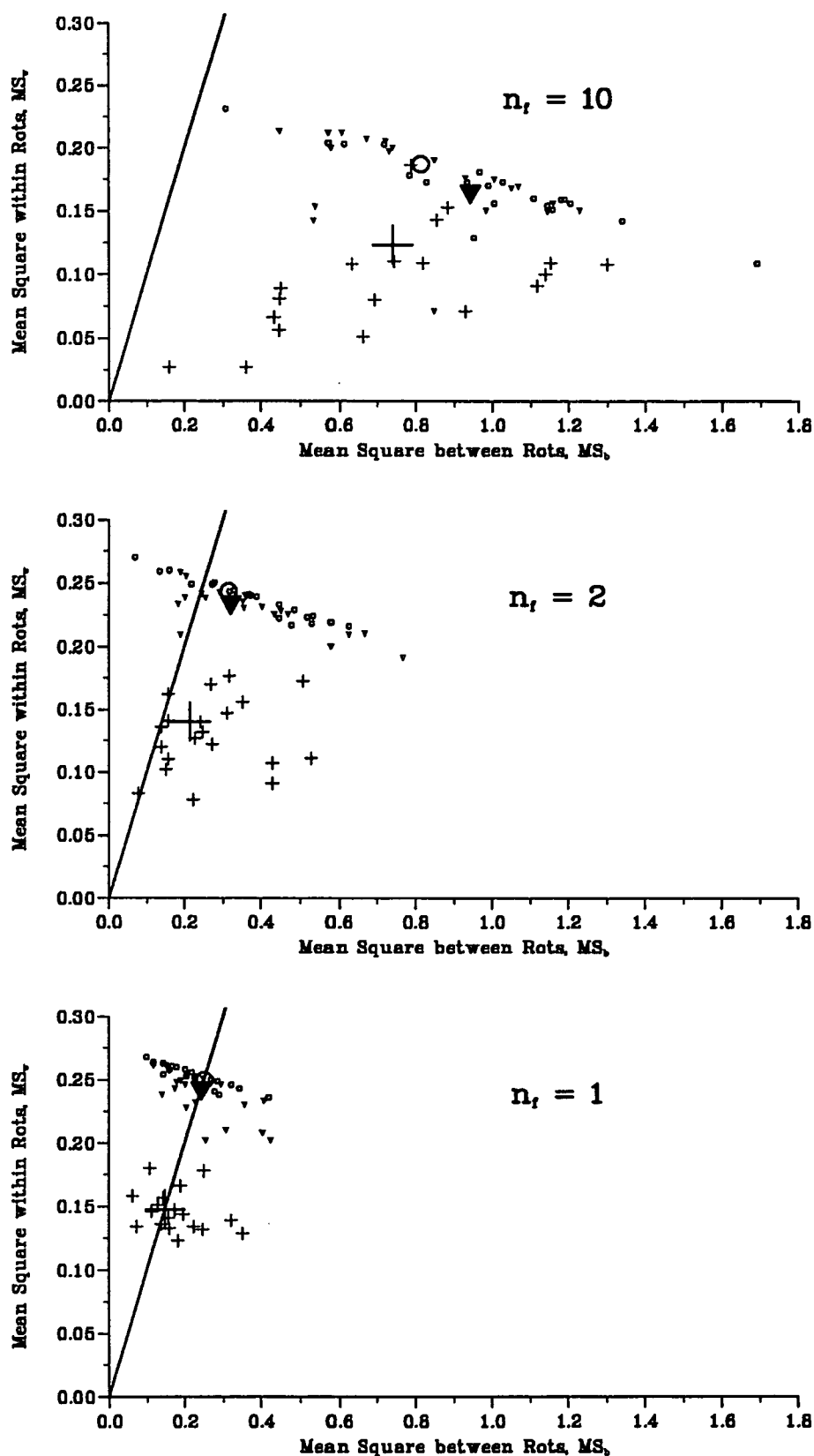


Figure 6.7: The observed MS_b and MS_w from an analysis of variance based on heterozygotes, for $n_f = 10, 2$ or 1 and $q = 0.5(\circ), 0.3(\nabla)$ and $0.1(+)$. The large symbol in each graph indicates the expected values and the line represents $MS_b = MS_w$.

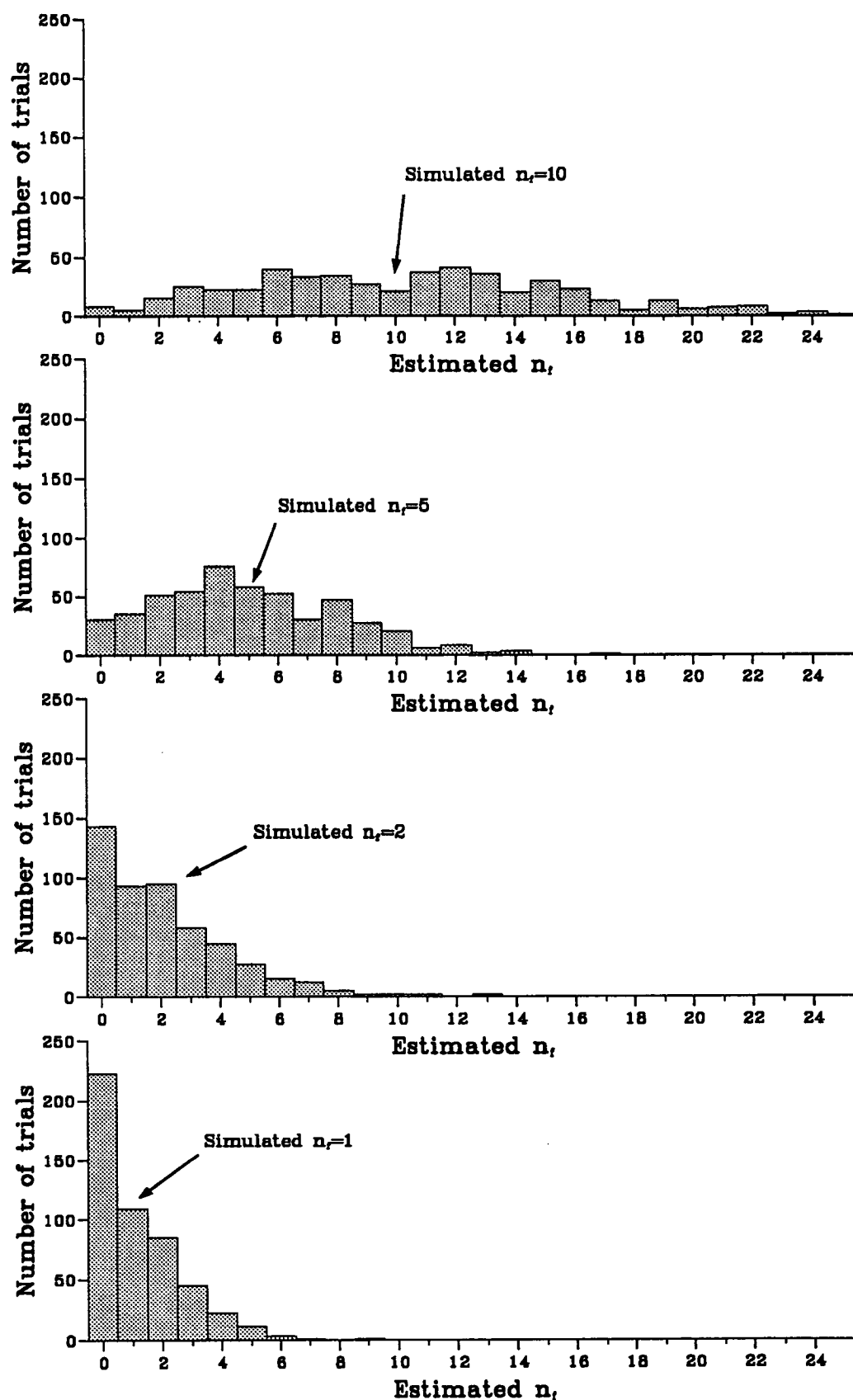


Figure 6.8: The estimates of family size, $\hat{n}_f$, for simulated $n_f = 1, 2, 5$ and 10, with $q = 0.5$. There are 500 replicates trials with $N = R = 10$ for each n_f , using data based on the number of heterozygotes.

replicate trials; here $q = 0.5$ and the simulated n_f is 1, 5 or 10 as indicated by the different symbols. With two equifrequent alleles, the estimate of n_f for the analysis based on heterozygotes is independent of q . The correlations between the estimates are 0.039, 0.148 and 0.576 for $n_f = 1, 5$ or 10, respectively. If the allele frequencies are not equal a different pattern of results emerges. Figure 6.9b shows results obtained from analysis of 100 replicate trials as before but with $q = 0.1$. Here, the correlations are 0.788, 0.872 and 0.807 for $n_f = 1, 5$ or 10, respectively. If the estimate of family size using genotype scores is denoted n_1 and the estimate based on heterozygotes is denoted n_2 , then $n_1 = n_f + e_1$, and $n_2 = n_f + e_2$ where e denotes the error. Using the method of generalised least squares, the best estimate of $\hat{n}_f$ is found to be

$$\hat{n}_f = \frac{[\text{Var}(n_1) - \text{Cov}(n_1, n_2)] n_2 + [\text{Var}(n_2) - \text{Cov}(n_1, n_2)] n_1}{\text{Var}(n_1) + \text{Var}(n_2) - 2\text{Cov}(n_1, n_2)},$$

and the variance of $\hat{n}_f$ is given by

$$\text{Var}(\hat{n}_f) = \frac{\text{Var}(n_1)\text{Var}(n_2) - (\text{Cov}(n_1, n_2))^2}{\text{Var}(n_1) + \text{Var}(n_2) - 2\text{Cov}(n_1, n_2)},$$

Using the results obtained from simulation experiments (as shown in Figures 6.9a and 6.9b), the observed covariance of n_1 and n_2 can be used to find approximate values for the combined $\hat{n}_f$. Taking the estimates of n_f from 200 simulations, the variances and covariances were found and used to calculate $\hat{n}_f$. The mean value of the combined $\hat{n}_f$, and the observed variances are given below.

q	n_f	Mean		Mean		Cov(n_1, n_2)		
		n_1	Var(n_1)	n_2	Var(n_2)		$\hat{n}_f$	Var($\hat{n}_f$)
0.1	1	0.894	0.708	0.882	2.755	1.100	0.898	0.586
	5	4.654	4.024	7.450	14.806	6.730	3.245	2.660
	10	9.527	7.221	16.413	23.870	10.591	7.185	6.076
0.5	1	0.932	0.668	0.914	2.459	0.050	0.929	0.542
	5	4.802	4.434	4.667	11.610	1.061	4.769	3.618
	10	9.430	7.505	9.258	25.610	7.991	9.435	7.491

Clearly, as $\text{Var}(n_2)$ is large, $\text{Var}(n_2) - \text{Cov}(n_1, n_2)$ is greater than $\text{Var}(n_1) - \text{Cov}(n_1, n_2)$ and the estimate from genotype scores (n_1) has a larger influence on the combined estimate. From these simulation results, when $q = 0.5$ the mean value of the combined estimate of n_f is very similar to the individual estimates, but the variance is smaller. When $q = 0.1$, the mean $\hat{n}_f$ is consistently lower than the true value. The variance of the combined estimate is also smaller than for either separate analysis.

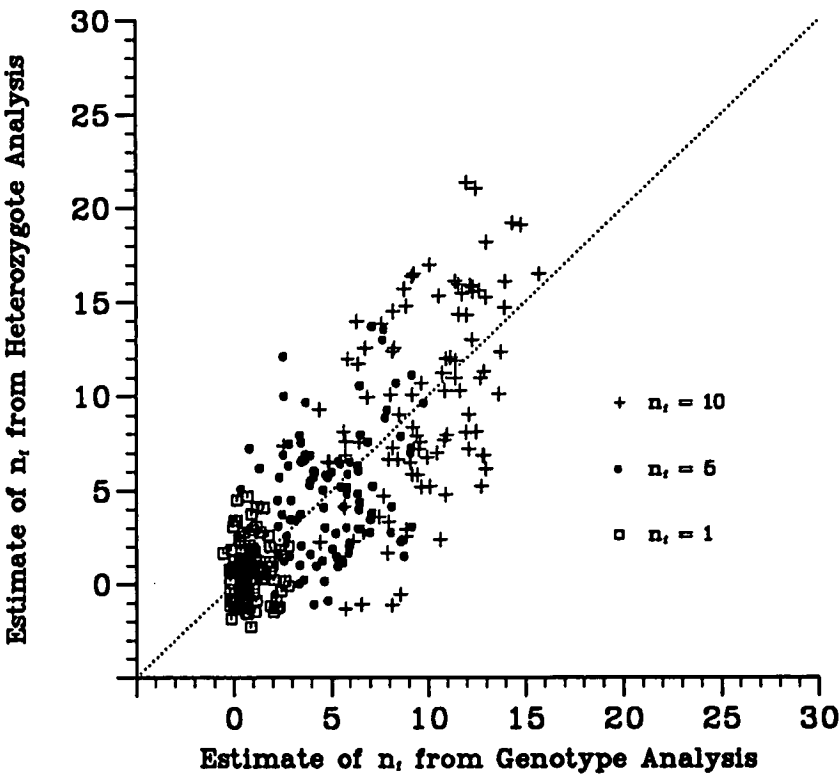


Figure 6.9a: The estimates of family size, $\hat{n}_i$, for simulated $n_i = 1, 5$ and 10, with $q = 0.5$. There are 100 replicates trials with $N = R = 10$ for each n_i , using data based on genotypes or the number of heterozygotes.

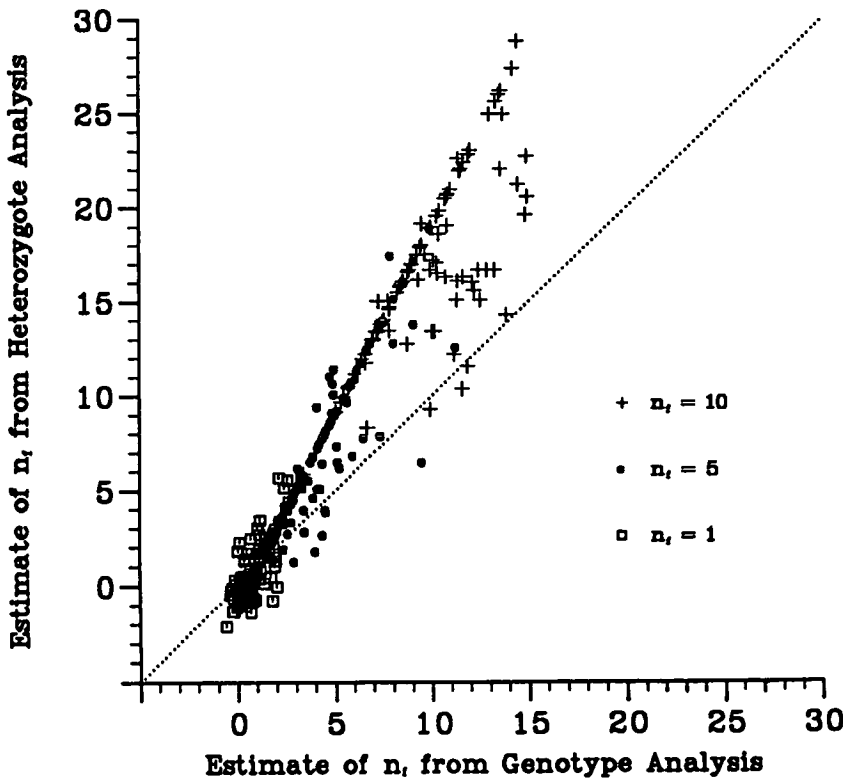


Figure 6.9b: As above, but with $q = 0.1$.

Note

The analysis of variance presented in Section 6.3 is based on the analysis of scores where the indicator variable x_{kij} equals 1 if the gene is A_1 and 0 otherwise, where $k = 1, 2, \dots, R$, $i = 1, 2, \dots, N$ and $j = 1, 2$. Table 6.1 shows that the total variation is due to differences between rots and differences within rots.

Table 6.8a shows an analysis of variance with three sources of variation. Here, the within rot component is divided into a *within individual* component and a *between individual* component. The within individual component is a measure of inbreeding and contains no information in this analysis.

Source of Variation	Degrees of Freedom	Sum of Squares
between rots	$R - 1$	$2N \sum_{k=1}^R (\bar{x}_{k..} - \bar{x}_{...})^2$
individuals within rots	$R(N - 1)$	$\sum_{k=1}^R \sum_{i=1}^N (x_{ki.} - \bar{x}_{k..})^2$
alleles in individuals	RN	$\sum_{j=1}^2 (x_{kij} - \bar{x}_{ki.})^2$
Total	$2RN - 1$	$\sum_{k=1}^R \sum_{i=1}^N \sum_{j=1}^2 (x_{kij} - \bar{x}_{...})^2$

Table 6.8a: The analysis of variance table showing three sources of variation.

The analysis presented in Section 6.4 is also based on diploid data but here the indicator variable $x_{ij} = 0$ if the genotype is A_1A_1 , 1 if the genotype is A_1A_2 and 2 if the genotype is A_2A_2 , where $i = 1, 2, \dots, R$ and $j = 1, 2, \dots, N$. In this case there are only N scores associated with each rot, each observation is essentially the sum of the two observations per individual used in the previous analysis (Section 6.3). The within rot variation cannot be further subdivided using data on genotypes.

6.7 Summary

In the first part of this chapter, a method was developed to estimate the mean family size in a model population from analysis of the genotypes of a large number of individuals.

In Section 6.3, the indicator variable x_{kij} took the value 1 if the gene j from an individual i from rot k was of type A_1 , and 0 otherwise. An analysis of the variance of these variables was used to estimate the family size within the sample. Simulation experiments show that the estimator $\hat{n}_j$ has a mean value close to the true n_j , but the variance is large, especially when n_j is large; when all the individuals sampled from each rot are sibs (one female laid N eggs on each rot) the variance of $\hat{n}_j$ is proportional to N^2 . When n_j is large the underlying variables x_{kij} are not normally distributed, so the expressions for the expected values are not accurate (also, the ratio of the mean squares does not necessarily have an F distribution). Similar empirical results were obtained with different allele frequencies.

In Section 6.4, the indicator variable x_{ij} took the value 0 if the genotype of individual j from rot i had genotype A_1A_1 , value 1 if the genotype was A_1A_2 and value 2 if the genotype was A_2A_2 . An analysis of variance of these variables produced very similar results to the analysis using allele scores. Taken over many replicate simulation trials, the mean estimate of family size was very close to the true value, but the variance was large. When the sample from a rot consists of one family of N sibs, the variance of $\hat{n}_j$ is proportional to N^2 , as before.

Finally, in Section 6.5, a third type of analysis was investigated where the indicator variable x_{ij} took the value 1 if the genotype of individual j from rot i had genotype A_1A_2 , but took the value 0 otherwise. This proved to be the least accurate method; estimates of n_j tended to be smaller than the true value, and the variance of $\hat{n}_j$ was larger than in the previous two examples.

The most informative analysis is the method based on genotype scores: the variance of $\hat{n}_j$ is smaller than with the other data types. This scoring method was therefore chosen for use with the field data described in Section 6.8.

6.8 Illustrative Example of Family Size Estimation Applied to Real Data

Drosophila species are the subject of many studies concerned with genetic variation, but experiments tend to be carried out using laboratory stocks. One exception is *Drosophila buzzatii*; these flies feed and breed exclusively on the cactus *Opuntia* and natural populations can be studied more easily than other populations of *Drosophila* in the wild. Thomas and Barker [1990] address the question of breeding structure in *D. buzzatii* by calculating measures of inbreeding and linkage disequilibrium. They are interested in determining the effective number of parents that produced the offspring emerging from the cactus rots. Results from the analysis carried out by these authors suggest that sib-groups are present in the flies emerging from a rot. One conclusion from the analysis, using the observed values of F_{ST} , is that approximately ten individuals contribute to the progeny emerging from a rot, i.e. about 5 parent-pairs. The F -statistics, F_{ST} , F_{IT} and F_{IS} , were introduced by Wright [1951] as a means of summarising population structure in subdivided natural populations. Similar parameters (θ , F and f , respectively) were used by Cockerham [1969]. Thomas and Barker [1990] estimated F_{ST} (the correlation between random gametes within population subdivisions relative to gametes of the total population), for the population of *D. buzzatii* using the methods of Weir and Cockerham [1984].

In this section, data supplied by R.H. Thomas are analysed using the method described in Section 6.4, and the results compared with those of Thomas and Barker. Collections of flies were made from individual rots from four sites in southeast Queensland (for details, see Thomas and Barker [1990]). Each rot was covered with gauze and flies emerging from each rot were aspirated daily; only data from fresh rots were used for the analysis in this thesis. In total, genotype data were supplied for 1582 flies from the four sites, for seven autosomal loci. A total of 15 records were discarded as they were not complete for seven loci. The data were supplied in two forms; the full genotypes and a two-allele collapse for each locus. Locus 7, *Lap*, was not used in the final analysis as most rots were invariant.

6.8.1 Summary of the Data

Allele frequencies, p_i , are calculated for each site separately and for all sites combined. Frequencies are estimated as

$$p_i = \frac{1}{2n} \sum_{a=1}^n \sum_{b=1}^2 x_{ab},$$

where $x_{ab} = 1$ if allele b for individual a is type A_i and n is the total number of individuals. The number of rots scored and the total number of flies sampled are shown in Table 6.8a. The number of flies sampled from each rot is similar for rots within sites but not exactly equal; the mean value is used in the calculations. Estimated allele frequencies are shown in Table 6.8b.

The majority of the rots were found to be in Hardy-Weinberg equilibrium; those that show a significant deviation have an excess of homozygotes. For all but nine rots where deviation from Hardy-Weinberg equilibrium was found, however, there are very small classes for the genotypes making the χ^2 test unreliable. A total of 8 rots over all sites showed a significant ($p < 0.1$) excess of heterozygotes for locus 6 (*Adh-1*). These results agree with those of Thomas and Barker [1990].

6.8.2 Family Size Estimation

For the method developed in the first part of this chapter, genotypes are scored for a single locus with two alleles. For each locus given in Table 6.8b the most common allele is therefore designated ' A_1 ' and all other alleles are pooled and called ' A_2 '. This two allele collapse of the data is then scored as described in Section 6.4, and the analysis of variance carried out. The observed ratio of mean squares is denoted F_G , and the family size, n_f , is estimated for each of the four sites, for each locus separately, using Expression (6.49). The results from the analysis are presented in Table 6.9. The values of $\hat{n}_f$ found from each locus are estimates of the same n_f and can be combined for each site. This is achieved by finding the mean F_G and using this value to calculate $\hat{n}_f$.

Estimates of family size for sites 2 and 3 are found from a very small number of rots (6 and 5 respectively) with about 15–20 flies sampled per rot; results for site 3 are very variable with $\hat{n}_f$ ranging from -0.9 to 10.37 for different loci. Results from site 2 are much less variable with values of $\hat{n}_f$ ranging from 0.20 to 2.86 (locus 2 did not

Site	Number of Rots	Number of Flies	Mean No: Flies per Rot
1	26	574	22.1
2	6	89	14.8
3	5	104	20.8
4	38	800	21.0
Total	65	1567	

Table 6.8a: The number of rots and the total number of flies sampled for each of the four sites.

Locus	Allele <i>i</i>	Allele Frequencies				
		Site 1	Site 2	Site 3	Site 4	All
1 <i>Pgm</i>	1	0.0218	0.0056	0.0240	0.0331	0.0268
	2	0.9782	0.9944	0.9760	0.9663	0.9729
	3	–	–	–	0.0006	0.0003
2 <i>Aldox</i>	1	0.9625	1.0000	0.9760	0.9906	0.9799
	2	0.0375	–	0.0240	0.0094	0.0201
3 <i>Est-1</i>	1	0.1969	0.1910	0.2788	0.1844	0.1956
	2	0.7761	0.7809	0.7019	0.7844	0.7757
	3	0.0270	0.0281	0.0192	0.0313	0.0287
4 <i>Est-2</i>	1	0.4172	0.3764	0.3510	0.5344	0.4703
	2	0.3920	0.4438	0.4038	0.2481	0.3223
	3	0.1115	0.0787	0.1058	0.1706	0.1394
	4	0.0793	0.1011	0.1394	0.0469	0.0680
5 <i>Hex</i>	1	0.8885	0.8876	0.8750	0.9200	0.9036
	2	0.1089	0.1124	0.1250	0.0800	0.0954
	3	0.0017	–	–	–	0.0006
	4	0.0009	–	–	–	0.0003
6 <i>Adh-1*</i>	2	0.5392	0.5337	0.5817	0.8281	0.6892
	3	0.4608	0.4663	0.4183	0.1719	0.3108
7 <i>Lap</i>	1	0.0017	–	–	0.0731	0.0380
	2	0.9965	1.0000	1.0000	0.8788	0.9368
	3	0.0017	–	–	0.0481	0.0252

* note that there are only 2 alleles labelled ‘2’ and ‘3’.

Table 6.8b: The allele frequencies calculated per site and for the total number of *Drosophila buzzatii* sampled.

Locus	Site							
	1		2		3		4	
	F_G	$\hat{n}_f$	F_G	$\hat{n}_f$	F_G	$\hat{n}_f$	F_G	$\hat{n}_f$
1 <i>Pgm</i>	3.68	5.56	0.63	0.29	0.02	-0.90	2.68	3.96
2 <i>Aldox</i>	3.23	4.87	—	—	2.39	3.48	2.05	2.91
3 <i>Est-1</i>	2.86	4.27	2.05	2.82	0.57	0.16	3.23	4.84
4 <i>Est-2</i>	3.13	4.71	1.58	2.04	7.45	10.37	2.06	2.92
5 <i>Hex</i>	2.20	3.17	1.86	2.52	6.42	9.19	2.25	3.25
6 <i>Adh-1</i>	2.79	4.16	2.07	2.86	3.13	4.68	2.70	4.00
pooled	2.98	4.47	1.64	2.14	3.33	4.99	2.50	3.67

Table 6.9: The observed ratio of the mean squares within and between rots, F_G , and the estimated family size, $\hat{n}_f$, for each of the sites.

vary at this site and so the analysis was not carried out). The pooled estimates over all loci for sites 2 and 3 are 2.14 and 4.99, respectively. Sites 1 and 4, however, include a moderately large number of rots (26 and 38, respectively) with 21 or 22 flies sampled per rot. For these examples, the estimates for $\hat{n}_f$ are consistent across loci. At site 1, results from each locus suggest that about 4–5 parent-pairs contributed to the progeny emerging from each rot. At site 4, results suggest 3–4 parent-pairs contributed to the progeny emerging from each rot. The pooled estimates over all loci for sites 1 and 4 are 4.47 and 3.67, respectively. Thomas and Barker [1990] also conclude that the number of parent-pairs is around five.

Hoffmann and Nielsen [1985] carry out a similar investigation into breeding structure for *Drosophila melanogaster*. The model population is considered to be subdivided into a number of ephemeral breeding sites; here, however, the number of offspring produced per site (e.g. fruit or rot) is assumed to depend on the genetic variance within the site. The relative fitnesses of the genotypes within a site are the same, but differ relative to other sites. Simulated data are generated and analysed to find expected values for Wrights F_{ST} statistic. Hoffmann and Nielsen captured flies as they emerged from rotting apples collected from the field and found the genotypes for three loci; from these data, values of F_{ST} are calculated. Comparison of the observed and expected

F_{ST} suggest that about 2–3 parent-pairs produced the offspring at each breeding site (i.e. each apple).

The values of family size calculated by the analysis of variance method are rough estimates; a number of the assumptions made in the model will not hold for a real population. For example, the assumption that equal numbers of females colonise each rot and produce equal numbers of offspring is unrealistic. The results obtained from analysis of the real data, however, show some consistency across loci and seem to be in agreement with work by other workers [Hoffmann & Nielsen, 1985; Thomas & Barker, 1990]. An analysis of field data may be further complicated by the fact that some cactus rots remain as suitable substrates for breeding for more than one generation; flies hatching together may breed and lay eggs in the same rot. Thomas and Barker [1990] suggests that a cactus rot may remain viable for as many as three generations, and so there will be inbreeding within the rot. This would tend to increase the estimate of family size by decreasing the variance within rots while increasing the variance between rots. The colonising females may also move between rots and there may be non-random selection of rots. Two major violations of the assumptions made in the model for natural populations are multiple mating by both parents and variable fitness in both parents and offspring. Including these factors in the model would require a more complex analysis procedure.

F -statistics are often used as a means of summarising population structure in subdivided natural populations. F (or F_{IT}) is the correlation between gametes that unite to produce the individuals, relative to the gametes of the total population (inbreeding) and θ (or F_{ST}) is the correlation between random gametes within subdivisions, relative to the gametes of the total population (coancestry). These parameters can be estimated from an analysis of variance of the frequencies of the allele of interest. Using the notation of Weir & Cockerham [1984], the expected values for the observed components of variance a for between populations, b for between individuals within populations and c for between gametes are

$$Ea = p(1 - p)\theta,$$

$$Eb = p(1 - p)(F - \theta) \text{ and}$$

$$Ec = p(1 - p)(1 - F)$$

where p is the frequency of the allele in the ancestral population. For the analysis presented in this thesis, variance among populations is assumed to be due to a single founding event, there is no inbreeding and the parents (founders) mate at random. The between rot variance, σ_b^2 , was found to be

$$q(1 - q) \frac{(n_f - 1)}{(N - 1)}$$

where N is the number of individuals (Expression (6.45)). If n_f is 1, $\sigma_b^2 = 0$ and as n_f increases, the between rot variance increases and the within rot variance decreases. In terms of θ , the between rot variance is $q(1 - q)\theta$. Thus there is a simple relationship between θ and n_f under the model described. When all samples are of the same size, Weir & Cockerham find that an unbiased estimator for θ is

$$\hat{\theta} = \frac{s^2 - \frac{1}{n-1} [\bar{p}(1 - \bar{p}) - \frac{r-1}{r} s^2 - \frac{1}{4} \bar{h}]}{\bar{p}(1 - \bar{p} + s^2/r)}$$

where r is the number of populations, s^2 is the sample variance of allele frequency over populations, $\bar{h}$ is the average heterozygote frequency for the allele of interest and $\bar{p}$ is the average sample frequency of the allele. This estimate of θ could be used to find an estimate of family size, n_f .

Chapter 7

Conclusions

7.1 Summary

Investigations into many aspects of population biology require an accurate assessment of genetic relationships. In particular, the detailed structure of a population can be inferred by examining the pattern of genealogical relationships between individual organisms. The inference of genetic relationship has been studied by several authors for a wide range of applications, but the approaches taken usually require data to be supplied from two or more generations. There are situations, however, where it is either impossible or impractical to obtain data from more than one generation.

The principle question posed at the beginning of this work was whether it is possible to infer the detailed, recent genealogy of a population by examining the genotypes observed in a sample drawn from a single generation, without knowledge of the immediate ancestors.

To answer this question, two different approaches are developed. The first approach aims to identify the detailed relationships between a small number of named individuals. The second approach aims to find an estimate of the relatedness among a large group of individuals where precise relationships are not important.

7.1.1 The Likelihood Method

This method is based on the probability of obtaining the genotypes observed in the sample given that the hypothetical relationship under consideration is the true one; this is termed the likelihood for the relationship. Details of the types of relationship structures included in the analysis are described in Chapter 2. A set of 14 family patterns are given, showing the possible genotypes that may be found in a set of full sibs. The likelihood for each of these families is calculated and the method is described for combining these expressions for more complex relationship structures consisting of many families. The relationship in which all individuals are considered as unrelated is chosen as the one with which all other relationships are compared, in the form of a likelihood ratio. The procedure for ranking the relationship hypotheses is also given.

The relationship structure that has the highest likelihood ratio is an estimate of the relatedness of the individuals in the sample. In Chapter 3, inferences of genetic relationship are based solely on the relationship ranked in first position; this relationship is taken as the best estimate of the relationships in the sample.

For a sample of two individuals, there are only two possible relationships and hence inference of relationship depends on whether the likelihood ratio is greater or less than one. Using data from a single locus, it is found that two individuals are inferred as sibs if they have identical genotypes and are inferred as unrelated if they have no alleles in common. This is true for any allele frequencies. If, however, the individuals have one allele in common, the inference of relationship depends on the frequency of the shared allele. For example, if there are only two alleles at the locus and the frequency of the shared allele is *less* than $1/3$, the individuals are inferred as sibs. For larger samples, the inference of relationship depends on the frequency of shared alleles and is shown to be complex, as switches in inference occur at several allele frequencies. Generally, individuals with identical genotypes are inferred as sibs whereas those with no alleles in common are inferred as unrelated. If a particular allele is rare, then individuals with copies of the allele tend to be inferred as sibs; this is intuitively reasonable as rare alleles are more likely to be identical by descent. The likelihood for relationships in which sibs share rare alleles can be orders of magnitude greater than the likelihoods for other relationships. When genotype data from several independent loci are included

in the analysis, the overall likelihood ratio for all loci combined is the product of the likelihood ratios for each locus separately. The combined likelihoods are found to be strongly influenced by rare alleles at any of the loci. For samples larger than two with data from a single locus, the relationship in which the individuals are unrelated is never inferred unless the alleles have intermediate frequency.

The likelihood for any relationship is readily found for a given a set of individuals with known genotypes. The subsequent inference of relationship is straightforward. A measure of the reliability of the method is given by the probability of correct inference of relationship. This probability is found theoretically for samples of two and three, and is estimated empirically for larger or multilocus examples. Using simulated genotype data, the proportion of samples in which the true relationship is inferred is an estimate of the probability of correct inference.

For two individuals and a single locus, the probability of correct inference is found to be high for full sibs if the allele frequencies are extreme, but very low for unrelated individuals under the same conditions. The probability of correct inference is approximately 0.6 for both relationship R_1 and R_2 when there are two equifrequent alleles; but with three equifrequent alleles, this falls to about 0.5. With increasing numbers of equifrequent alleles, however, the probability of correct inference depends on the true relationship. For example, with five equifrequent alleles the probabilities of correct inference for R_1 and R_2 are 0.55 and 0.8, respectively. With a large number of rare alleles the probability of correct inference increases to 0.8 for both relationships. However, the probability of correct inference will always be less than one as a proportion of samples will be incorrectly inferred.

One unexpected result is that the number of samples correctly inferred for any sample size could *decrease* when data from two loci are used instead of one locus. The pattern of inference for each possible sample of two individuals is examined in detail: the inference of relationship may be independent of allele frequency or depend on the frequencies at one or both loci. Thus the overall probability of correct inference is not a simple function of allele frequency for two individuals. However, the probability of correct inference does increase when data from more than two loci are included. For a sample of two individuals, the probability of a correct inference of relationship is found to be consistently in the range 0.7–0.8 when typical field data are used. It

is found that the probability of correct inference decreases for all relationships as the sample size increases. Generally, this probability is higher when the individuals are all siblings. With six individuals in the sample, using field data, the probability of correct inference ranges from 0.5 down to 0.007 for full sibs and unrelated individuals, respectively. These figures were consistently obtained using different numbers of loci and alleles.

The analysis is extended in Chapter 4 to take account of the trends within the full list of ranked relationship structures, in contrast to Chapter 3 where only the first ranked relationship is considered. A more general, and in many ways more informative, picture can be built up by examining the pattern of relationship structures after sorting them with respect to the likelihood for each relationship.

For any sample with more than two individuals there are a variable number of relationships to consider, depending on the observed genotypes. Provided there are more than two alleles at any one locus some relationships are found to have a likelihood equal to zero; such relationships are *excluded*. The excluding power of a locus is found to be maximal when the alleles have equal frequency. Exclusion cannot unequivocally determine the true relationship between individuals because there will always be more than one valid relationship structure for a given set of individuals. However, useful information regarding relationships between members of the sample can be found by examining the type of excluded relationships. For example, the true relationship and those that contain only true subsets of families cannot be excluded (for families larger than two).

The relationship structures that are not excluded for a given sample are termed valid relationships. For samples in which the individuals are closely related, it has been shown that the true relationship and relationships that are similar to it will be placed in the higher ranks, while structures that split up true sibships will be placed in lower ranks. Alternatively, for samples in which the individuals are not closely related then the order of relationships in the ranked list does not show any consistent pattern.

Expected likelihood ratios are found for small samples and estimated from simulation experiments for larger samples. The true relationship does not necessarily give

the maximum expected likelihood ratio, and hence the method appears not to be consistent. Thompson [1976b] found that a method comparing *relationships* for given individuals was consistent but a method comparing *individuals* for given relationships was not. For genealogies spread over many generations it is not feasible to consider all possible genealogies and so the estimated structure is built up from smaller sub-units of parents and offspring (see also Section 1.2). However, the process of building up a genealogy from subunits involves comparing individuals for given relationships rather than relationships between a given set of individuals. When pairs of individuals are considered as parent and offspring, it is not necessarily the true parents who give maximal expected log-likelihood for the hypothesis; on average true siblings are chosen as parents provided they are not excluded. Thompson concludes that despite this paradox, inclusion of age data reduced this problem and accurate genealogies could still be reconstructed. For the analysis in this thesis, it is shown that when the individuals are all sibs, the true relationship has the maximum expected likelihood ratio. When the individuals are unrelated the expected likelihood ratios are all equal to one, because the likelihood ratios are all calculated relative to the true relationship. For other relationship structures, there are a number of families of different sizes depending on the sample size. For example, if there are three individuals then the relationship structure with two families has one family of size two and one family of size one. Within this relationship there are different ways to divide the individuals; thus different individuals are compared for the same relationship. The true relationship does not have maximum expected likelihood ratio in this case.

Single samples of three, four or five individuals are examined in Chapter 5 to demonstrate the interpretation of data using the methods described in the previous chapters. Four examples are simulated with known relationship, while four more samples are taken from a natural population in which the true relationship is not known. For two of the simulated samples, the true relationship is ranked first and so is correctly inferred; the pattern of relationships in the list of ranked likelihoods confirms the conclusions. For the third sample, the wrong conclusion is reached, demonstrating that the method is unreliable even with data from several loci. The last of the simulated samples highlights some of the problems associated with interpreting the results when there are many possible relationship structures. Several of the relationships are excluded, and

the pattern seen within the list of valid relationships suggests the individuals are not closely related; this conclusion is correct although the true relationship is not ranked first. For the second set of examples, the results obtained from analysis using the likelihood method agree well with those obtained by direct comparison with individuals from the parental generation. It is not possible, using the present model, to distinguish between half sibs and so the estimate of relationship is essentially a guide to the degree of relatedness in the sample.

7.1.2 The Analysis of Variance Method

A different approach is developed in Chapter 6. Here the aim is to estimate the number of parents that contributed to the observed offspring generation. This is achieved by analysis of the genotypes of a large number of unnamed individuals from the offspring generation. For the analysis, an indicator variable is defined and an analysis of variance of these variables carried out. Three types of data are used for comparison; allele counts, genotype counts and scores based on heterozygosity. The number of individuals per family within the offspring generation is estimated from the ratio of the between and within group mean squares; the estimator is found to be independent of allele frequency. From this, the number of parents is found based on a simple model population.

The most informative analysis proved to be the method based on genotype scores; the variance of the estimator is smaller than with the other methods and thus is more efficient. Simulation experiments show that the estimator for family size, $\hat{n}_f$, has a mean value close to the true n_f , but the variance is large. When n_f is large, results from simulation based on the simple model do not agree well with theoretical values. When the sample from a rot consists of one family of N sibs, the variance of $\hat{n}_f$ is proportional to the square of the sample size.

The results obtained when this method of analysis is applied to a set of field genetic data are given in Section 6.8. The estimates of family size agree with those of other workers, although sites with few data give very variable results.

7.2 Conclusions and Extensions

The results from both methods of analysis suggest that some information regarding genetic relationship can be obtained from the genotypes of individuals from a single generation.

If sufficient informative loci are available, then the relationship between a small number of individuals can be estimated. Using only the highest ranked hypothesis to estimate the genetic relationship between individuals is shown to be unreliable and the accuracy of the inference depends on the true relationship. The probability that a relationship will be correctly inferred is greater when the individuals are from one or a few families, but is low in most cases. Sibships with three or more full members may be excluded on the basis of incompatible genotypes, but when they are not excluded the likelihood ratios for these relationships tend to be high. In effect, the likelihood method picks out any *possible* relationships, so that closer relationships than actually exist are often inferred, particularly when some loci have extreme allele frequencies. This in turn decreases the probability of correct inference for relationships that have many, small families as these will rarely be inferred. Even with a large number of loci the probability of correct inference for an unrelated group is very much lower than was expected. The probability of correct inference decreases with increasing sample size, indicating that the analysis is not suitable for use with large samples.

A more promising approach involves examining the pattern of relationship structures after they have been ranked according to the magnitude of their likelihood ratios. Simulation results show that distinct patterns can be seen for different relationships. When a single sample of individuals is analysed, the relative order of relationship structures, together with information from exclusion, can give useful information about relatedness. With further work, it should be possible to develop a formal pattern recognition and interpretation procedure to be applied to single samples.

If it is not necessary to know the exact relationships between named members of the group, or samples are very large, then the analysis of variance method is used. The approach presented in this thesis is a preliminary analysis, based upon a simple model population. Results from simulations show that, as with the likelihood method, some

information regarding genetic relationship can be gained but the accuracy depends on the true relatedness. The main drawback with the basic model is that the family size is assumed to be constant for each female. A more realistic model would entail a variable number of parents colonising each new breeding site, with a variable number of offspring produced per parent.

There are several factors that limit the application of the likelihood method and should be addressed by future work.

- It has been assumed that the genotype of each individual is known. When there is dominance, only the phenotypes of the individuals may be observable. The probability of obtaining the observed phenotypes for given relationships could be calculated and a set of *phenotypic* family patterns found, similar to the *genotypic* patterns in Section 2.3.3. From this, the likelihood for each relationship could be calculated as before. A locus exhibiting dominance provides less information than if the genotypes were distinguishable and inferences of relationship would be less accurate.
- In natural populations, other relationships would exist in addition to the structures described in Section 2.2. The most common would be half sibships due to breeding patterns other than monogamy. In an experiment to estimate the frequency of multiple parentage, the female parent is usually known, particularly for mammals, so offspring are known to be either full or half sibs. The analysis is then essentially an investigation into multiple paternity, and if the females genotype is available an analysis that includes this extra information will be more powerful than one that does not. However, for situations where the female is unavailable, the likelihood method could be extended to include half sibs. The joint probabilities for an ordered set of s sibs *and* a parent are readily calculated for a single locus. There are four family patterns for s sibs with one homozygous parent and 17 family patterns of s sibs and a heterozygote parent, similar to the patterns described in Section 2.3.3. The drawback is that the total number of combinations of individuals into families including full and half sibs would be very large and the likelihood ratios for each structure would be difficult to interpret. The use of X-linked loci would also require a distinction to be made between male and female parents.

- Throughout Chapters 2 to 4, the likelihood for each relationship is calculated by substituting the appropriate values for the allele frequencies into the expressions listed in Section 2.3.3. The relationship for which the likelihood is highest is then found by comparison. The analysis has thus been restricted to the case where allele frequencies are known. If the population allele frequencies are not known an alternative procedure is required to find the relationship with maximum likelihood. For any sample, a set of relationship structures can be defined and the likelihood for each of these structures can be found in terms of the allele frequencies, p_i ; these likelihoods can be maximised. In Chapter 3, it was shown that for a sample of two individuals and a single locus with two alleles the relationship with maximum likelihood can be found for each possible pair of genotypes. For this simple example, the maximum point for each graph is clear (see Figure 3.1 and associated text). When there are more than two alleles at each of several loci, and more than two individuals, the analysis is more complex. However, the likelihoods could be maximised for each relationship and the highest of these values used to infer the relationship in the sample.

As the sample size increases, the number of possible relationships, and therefore the number of calculations required, increases rapidly. Results have shown that the probability of correct inference is reasonably high when there are two individuals in the sample. This suggests that an alternative approach could be based on the genotypes of the members of the sample compared in sets of two. Cluster analysis is a term used for techniques that seek to separate data into constituent groups. Such techniques could be used to identify “families” as groups of similar individuals. The majority of clustering techniques begin with the calculation of a matrix of similarities or distances between objects; the likelihood ratio for a pair of individuals could be used as such a measure. When grouping individuals, at one extreme there are N families each with one individual and at the other extreme there is one family of N sibs. Agglomerative techniques proceed by fusing similar individuals into groups whereas divisive techniques begin with one group and split this into sub-groups. Deciding on the optimum number of groups, and hence the inference of relatedness, would be an interesting task.

The analysis techniques described in this thesis have been applied to both simulated and field data for samples with a range of sizes and relatedness. When data are available

from only one generation, it has been shown that within the limitations of the models presented here, these analysis methods enable certain inferences to be made regarding genealogical relationship. In the long run, the true relationship is most frequently estimated by both methods. However, the degree of relatedness of the individuals is a factor in determining the accuracy with which the family groups can be inferred. With small numbers of individuals, the more closely related the sampled individuals, the higher the probability of correct inference. With large samples, the estimator for family size has greater variance when the sampled individuals are from large families.

In practise, the relatedness within the study population is not known *a priori*, so the analysis methods at present have limited use when applied in isolation. If used in conjunction with other techniques and experimental observations they may provide extra, valuable information.

Appendix A

The Likelihood Equations

In Chapter 2, the procedure for the calculation of the likelihood for a given genetic relationship is shown (Section 2.3.2). The concept of *family patterns* is introduced in Section 2.3.3; these patterns, denoted Δ_i , are sets of genotypes that could be observed in a family of full siblings. The probability of obtaining an ordered sample of n individuals with genotype pattern Δ_i can be found for each of these 14 types (from Expression (2.3)) for a single locus with s alleles. These expressions can be simplified and are the joint genotype distributions for n siblings.

This Appendix shows the complete calculation of the joint genotype distributions for each Δ_i given that the sample are a) all full siblings and b) all unrelated.

Family Pattern Δ_1 : A sample is pattern Δ_1 for a locus if all n members have genotype $A_i A_i$ at that locus.

Relationship $R_{1:n}$:

Parent Pair	$\Pr(\Delta_1 \mid \text{Parent Pair})$
1. $A_i A_i \times A_i A_i$	$1^n p_i^4$
2. $A_i A_i \times A_j A_j$	0
3. $A_i A_i \times A_i A_j$	$\sum_{\substack{j=1 \\ j \neq i}}^s \left(\frac{1}{2}\right)^n 4p_i^3 p_j$
4. $A_i A_i \times A_j A_k$	0
5. $A_i A_j \times A_i A_j$	$\sum_{\substack{j=1 \\ j \neq i}}^s \left(\frac{1}{4}\right)^n 4p_i^2 p_j^2$
6. $A_i A_j \times A_i A_k$	$\frac{1}{2} \sum_{\substack{j=1 \\ k \neq j \neq i}}^s \sum_{k=1}^s \left(\frac{1}{4}\right)^n 8p_i^2 p_j p_k$
7. $A_i A_j \times A_k A_l$	0

Note that the probabilities must be summed for all possible allele combinations, and that for the general parent pair type numbered 6 above, the probability must be divided by two to account for the different ordering of the parents (or specify $j < k$). Summing these probabilities for all possible parent pairs gives

$$\Pr(\Delta_1 \mid R_{1:n}) = \left(\frac{1}{4}\right)^{n-1} p_i^2 \left[4^{n-1} p_i^2 + 2^n p_i \sum_{\substack{j=1 \\ j \neq i}}^s p_j + \sum_{\substack{j=1 \\ j \neq i}}^s p_j^2 + \sum_{\substack{j=1 \\ k \neq j \neq i}}^s \sum_{k=1}^s p_j p_k \right]$$

But $\sum_{\substack{j=1 \\ j \neq i}}^s p_j = 1 - p_i$ and $\sum_{j=1}^s \sum_{k=1}^s p_j p_k = \left(\sum_{\substack{j=1 \\ j \neq i}}^s p_j\right)^2 + \sum_{\substack{j=1 \\ j \neq i}}^s p_j^2$ for $k \neq j \neq i$,

therefore,

$$\begin{aligned} \Pr(\Delta_1 \mid R_{1:n}) &= \left(\frac{1}{4}\right)^{n-1} p_i^2 \left[4^{n-1} p_i^2 + 2^n p_i (1 - p_i) + (1 - p_i)^2 \right] \\ &= \left(\frac{1}{4}\right)^{n-1} p_i^2 \left[p_i^2 (4^{n-1} - 2^n + 1) + 2p_i (2^{n-1} - 1) + 1 \right] \\ &= \left(\frac{1}{4}\right)^{n-1} p_i^2 \left[(2^{n-1} - 1) p_i + 1 \right]^2. \end{aligned}$$

Relationship $R_{n:1,1,\dots,1}$:

$$\Pr(\Delta_1 \mid R_{n:1,1,\dots,1}) = \prod_{x=1}^n \Pr(\gamma_x) = p_i^{2n}$$

where γ_x is the genotype of individual x in the sample.

Family Pattern Δ_2 : A Δ_2 family has n members with identical genotypes $A_i A_j$.

Relationship $R_{1:n}$;

Parent Pair	$\Pr(\Delta_2 \mid \text{Parent Pair})$	Example for $A_1 A_2$
1. $A_i A_i \times A_i A_i$	0	—
2. $A_i A_i \times A_j A_j$	$1^n 2p_i^2 p_j^2$	$A_1 A_1 \times A_2 A_2$
3. $A_i A_i \times A_i A_j$	$\left(\frac{1}{2}\right)^n 4p_i^3 p_j$ $+ \left(\frac{1}{2}\right)^n 4p_i p_j^3$	$A_1 A_1 \times A_1 A_2$ $A_2 A_2 \times A_1 A_2$
4. $A_i A_i \times A_j A_k$	$\sum_{\substack{k=1 \\ k \neq j \neq i}}^s \left(\frac{1}{2}\right)^n 4p_i^2 p_j p_k$ $+ \sum_{\substack{k=1 \\ k \neq j \neq i}}^s \left(\frac{1}{2}\right)^n 4p_i p_j^2 p_k$	$A_1 A_1 \times A_2 A_k$ $A_2 A_2 \times A_1 A_k$
5. $A_i A_j \times A_i A_j$	$\left(\frac{1}{2}\right)^n 4p_i^2 p_j^2$	$A_1 A_2 \times A_1 A_2$
6. $A_i A_j \times A_i A_k$	$\sum_{\substack{k=1 \\ k \neq j \neq i}}^s \left(\frac{1}{4}\right)^n 8p_i^2 p_j p_k$ $+ \sum_{\substack{k=1 \\ k \neq j \neq i}}^s \left(\frac{1}{4}\right)^n 8p_i p_j^2 p_k$ $+ \sum_{\substack{k=1 \\ k \neq j \neq i}}^s \left(\frac{1}{4}\right)^n 8p_i p_j p_k^2$	$A_1 A_2 \times A_1 A_k$ $A_1 A_2 \times A_2 A_k$ $A_1 A_k \times A_2 A_k$
7. $A_i A_j \times A_k A_l$	$\sum_{\substack{k=1 \\ k \neq j}}^s \sum_{\substack{l=1 \\ l \neq k \neq j}}^s \left(\frac{1}{4}\right)^n 8p_i p_j p_k p_l$	$A_1 A_k \times A_2 A_l$

So, summing the probabilities for all possible parents,

$$\begin{aligned}
 \Pr(\Delta_2 \mid R_{1:n}) &= \left(\frac{1}{2}\right)^n p_i p_j \left[2^{n+1} p_i p_j + 4p_j^2 + 4p_i p_j\right] \\
 &\quad + \left(\frac{1}{4}\right)^n p_i p_j \sum_{\substack{k=1 \\ k \neq j \neq i}}^s p_k \left[2^{n+2} p_i + 2^{n+2} p_j + 8p_i + 8p_j\right] \\
 &\quad + \left(\frac{1}{4}\right)^n 8p_i p_j \left[\sum_{\substack{k=1 \\ k \neq j \neq i}}^s p_k^2 + \sum_{\substack{k=1 \\ l \neq k \neq j}}^s \sum_{\substack{l=1 \\ l \neq k \neq j}}^s p_k p_l \right]
 \end{aligned}$$

But, $\sum_{k=1}^s \sum_{l=1}^s p_k p_l$ for $l \neq k \neq j \neq i$ is equal to $(\sum_{k=1}^s p_k)^2 - \sum_{k=1}^s p_k^2$, $k \neq j \neq i$ and $\sum_{\substack{k=1 \\ k \neq j \neq i}}^s p_k = 1 - p_i - p_j$. Therefore,

$$\Pr(\Delta_2 \mid R_{1:n}) = \left(\frac{1}{4}\right)^{n-1} p_i p_j \left[2^n 2^{n-1} p_i p_j + 2^n p_i^2 + 2^n p_j^2 + 2^n p_i p_j\right]$$

$$\begin{aligned}
& + (1 - p_i - p_j)(2^n p_i + 2^n p_j + 2p_i + 2p_j) + 2(1 - p_i - p_j)^2 \Big] \\
& = 2 \left(\frac{1}{4} \right)^{n-1} p_i p_j \left[2^{n-1}(2^{n-1} - 1)p_i p_j + (2^{n-1} - 1)p_i + (2^{n-1} - 1)p_j + 1 \right].
\end{aligned}$$

Relationship $R_{n:1,1,\dots,i}$

$$\Pr(\Delta_2 \mid R_{n:1,1,\dots,i}) = \prod_{x=1}^n \Pr(\gamma_x) = (2p_i p_j)^n.$$

Family Pattern Δ_3 : A family of n individuals in which r_1 have genotype $A_i A_i$ and r_2 have genotype $A_j A_j$ will be classified as pattern Δ_3 ($r_1 + r_2 = n$ and $j \neq i$).

Relationship $R_{1:n}$

Parent Pair	$\Pr(\Delta_3 \mid \text{Parent Pair})$
1. $A_i A_i \times A_i A_i$	0
2. $A_i A_i \times A_j A_j$	0
3. $A_i A_i \times A_i A_j$	0
4. $A_i A_i \times A_j A_k$	0
5. $A_i A_j \times A_i A_j$	$\left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} 4p_i^2 p_j^2$
6. $A_i A_j \times A_i A_k$	0
7. $A_i A_j \times A_k A_l$	0

Only one general type of parental cross can produce a family with both $A_i A_i$ and $A_j A_j$ offspring, so

$$\Pr(\Delta_3 \mid R_{1:n}) = \left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} 4p_i^2 p_j^2 = \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j^2.$$

Relationship $R_{n:1,1,\dots,i}$

$$\Pr(\Delta_3 \mid R_{n:1,1,\dots,i}) = \prod_{x=1}^n \Pr(\gamma_x) = p_i^{2r_1} p_j^{2r_2}.$$

Note that in each case the family of n individuals are considered as ordered.

Family Pattern Δ_4 : A Δ_4 pattern is characterised by having r_1 individuals with genotype $A_i A_i$ and r_2 with genotype $A_i A_j$ only.

Relationship $R_{1:n}$:

Parent Pair	$\Pr(\Delta_4 \mid \text{Parent Pair})$
1. $A_i A_i \times A_i A_i$	0
2. $A_i A_i \times A_j A_j$	0
3. $A_i A_i \times A_i A_j$	$(\frac{1}{2})^{r_1} (\frac{1}{2})^{r_2} 4p_i^3 p_j$
4. $A_i A_i \times A_j A_k$	0
5. $A_i A_j \times A_i A_j$	$(\frac{1}{4})^{r_1} (\frac{1}{2})^{r_2} 4p_i^2 p_j^2$
6. $A_i A_j \times A_i A_k$	$\sum_{\substack{k=1 \\ k \neq j \neq i}}^s (\frac{1}{4})^{r_1} (\frac{1}{4})^{r_2} 8p_i^2 p_j p_k$
7. $A_i A_j \times A_k A_l$	0

Noting that $\sum_{\substack{k=1 \\ k \neq j \neq i}}^s p_k = 1 - p_i - p_j$, summing for all possible parent pairs we have

$$\begin{aligned}
 \Pr(\Delta_4 \mid R_{1:n}) &= \left(\frac{1}{2}\right)^{r_1+r_2} 4p_i^3 p_j + \left(\frac{1}{4}\right)^{r_1+r_2-1} 2^{r_2} p_i^2 p_j^2 + \left(\frac{1}{2}\right)^{r_1+r_2-1} 2p_i^2 p_j (1 - p_i p_j) \\
 &= \left(\frac{1}{4}\right)^{r_1+r_2-1} p_i^2 p_j \left[2^{r_1+r_2} p_i + 2^{r_2} p_j + 2 - 2p_i - 2p_j\right] \\
 &= 2 \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j \left[(2^{n-1} - 1)p_i + (2^{r_2-1} - 1)p_j + 1\right].
 \end{aligned}$$

Relationship $R_{n:1,1,\dots}$:

$$\Pr(\Delta_4 \mid R_{n:1,1,\dots}) = \prod_{x=1}^n \Pr(\gamma_x) = p_i^{2r_1} (2p_i p_j)^{r_2} = 2^{r_2} p_i^{2r_1+r_2} p_j^{r_2}.$$

Family Pattern Δ_5 : There are r_1 individuals with genotype $A_i A_i$, r_2 with genotype $A_i A_j$ and r_3 with genotype $A_j A_j$ in a family of type Δ_5 .

Relationship $R_{1:n}$: These individuals can only be full sibs if the parents are both $A_i A_j$,

Parent Pair	$\Pr(\Delta_5 \mid \text{Parent Pair})$
1. $A_i A_i \times A_i A_i$	0
2. $A_i A_i \times A_j A_j$	0
3. $A_i A_i \times A_i A_j$	0

4. $A_i A_i \times A_j A_k$ 0
5. $A_i A_j \times A_i A_j$ $\left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{2}\right)^{r_2} \left(\frac{1}{4}\right)^{r_3} 4p_i^2 p_j^2$
6. $A_i A_j \times A_i A_k$ 0
7. $A_i A_j \times A_k A_l$ 0

$$\Pr(\Delta_5 \mid R_{1:n}) = \left(\frac{1}{4}\right)^{r_1+r_3-1} \left(\frac{1}{2}\right)^{r_2} p_i^2 p_j^2 = 2^{r_2} \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j^2.$$

Relationship $R_{n:1,1,\dots}$;

$$\Pr(\Delta_5 \mid R_{n:1,1,\dots}) = \prod_{x=1}^n \Pr(\gamma_x) = p_i^{2r_1} (2p_i p_j)^{r_2} p_j^{2r_3} = 2^{r_2} p_i^{2r_1+r_2} p_j^{r_2+2r_3}.$$

Family Pattern Δ_6 : Here r_1 of the offspring are $A_i A_i$ and r_2 are $A_j A_k$.

Relationship $R_{1:n}$; Again, only one parental cross can produce a family with these genotypes.

Parent Pair	$\Pr(\Delta_6 \mid \text{Parent Pair})$
1. $A_i A_i \times A_i A_i$	0
2. $A_i A_i \times A_j A_j$	0
3. $A_i A_i \times A_i A_j$	0
4. $A_i A_i \times A_j A_k$	0
5. $A_i A_j \times A_i A_j$	0
6. $A_i A_j \times A_i A_k$	$\left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} 8p_i^2 p_j p_k$
7. $A_i A_j \times A_k A_l$	0

$$\Pr(\Delta_6 \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{r_1+r_2-1} p_i^2 p_j p_k = 2 \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j p_k.$$

Relationship $R_{n:1,1,\dots}$;

$$\Pr(\Delta_6 \mid R_{n:1,1,\dots}) = \prod_{x=1}^n \Pr(\gamma_x) = p_i^{2r_1} (2p_j p_k)^{r_2}.$$

Family Pattern Δ_7 : A Δ_7 pattern sample has r_1 individuals with genotype $A_i A_j$ and r_2 with genotype $A_i A_k$.

Relationship $R_{1:n}$;

Parent Pair	$\Pr(\Delta_7 \mid \text{Parent Pair})$	Example for $A_1 A_2$ and $A_1 A_3$ offspring
1. $A_i A_i \times A_i A_i$	0	—
2. $A_i A_i \times A_j A_j$	0	—
3. $A_i A_i \times A_i A_j$	0	—
4. $A_i A_i \times A_j A_k$	$\left(\frac{1}{2}\right)^{r_1} \left(\frac{1}{2}\right)^{r_2} 4p_i^2 p_j p_k$	$A_1 A_1 \times A_2 A_3$
5. $A_i A_j \times A_i A_j$	0	—
6. $A_i A_j \times A_i A_k$	$\left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} 8p_i^2 p_j p_k$ $+\left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} 8p_i p_j^2 p_k$ $+\left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} 8p_i p_j p_k^2$	$A_1 A_2 \times A_1 A_3$ $A_1 A_2 \times A_2 A_3$ $A_1 A_3 \times A_2 A_3$
7. $A_i A_j \times A_k A_l$	$\sum_{\substack{l=1 \\ l \neq k \neq j \neq i}}^s \left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} 8p_i p_j p_k p_l$	$A_1 A_l \times A_2 A_3$

Summing over all possible parent pairs, and noting that $\sum_{l=1}^s p_l = (1 - p_i - p_j - p_k)$ for $l \neq k \neq j \neq i$,

$$\begin{aligned}
 \Pr(\Delta_7 \mid R_{1:n}) &= \left(\frac{1}{2}\right)^{r_1+r_2} \left(\frac{1}{4}\right)^{-1} p_i^2 p_j p_k + \left(\frac{1}{4}\right)^{r_1+r_2-1} 2p_i^2 p_j p_k \\
 &\quad + \left(\frac{1}{4}\right)^{r_1+r_2-1} 2p_i p_j^2 p_k + \left(\frac{1}{4}\right)^{r_1+r_2-1} 2p_i p_j p_k^2 \\
 &\quad + \left(\frac{1}{4}\right)^{r_1+r_2-1} 2p_i p_j p_k \sum_{\substack{l=1 \\ l \neq k \neq j \neq i}}^s p_l \\
 &= \left(\frac{1}{4}\right)^{r_1+r_2-1} p_i p_j p_k \left[2^{r_1+r_2} p_i + 2p_i + 2p_j + 2p_k + 2(1 - p_i - p_j - p_k) \right] \\
 &= 2 \left(\frac{1}{4}\right)^{r_1+r_2-1} p_i p_j p_k \left[2^{r_1+r_2-1} p_i + 1 \right] = 2 \left(\frac{1}{4}\right)^{n-1} p_i p_j p_k \left[2^{n-1} p_i + 1 \right].
 \end{aligned}$$

Relationship $R_{n:1,1,\dots}$;

$$\Pr(\Delta_7 \mid R_{n:1,1,\dots}) = \prod_{x=1}^n \Pr(\gamma_x) = (2p_i p_j)^{r_1} (2p_i p_k)^{r_2} = 2^n p_i^n p_j^{r_1} p_k^{r_2}.$$

Family Pattern Δ_8 : Pattern Δ_8 has r_1 individuals with genotype $A_i A_i$, r_2 with genotype $A_i A_j$ and r_3 individuals with genotype $A_i A_k$.

Relationship $R_{1:n}$

Parent Pair	$\Pr(\Delta_8 \mid \text{Parent Pair})$
1. $A_i A_i \times A_i A_i$	0
2. $A_i A_i \times A_j A_j$	0
3. $A_i A_i \times A_i A_j$	0
4. $A_i A_i \times A_j A_k$	0
5. $A_i A_j \times A_i A_j$	0
6. $A_i A_j \times A_i A_k$	$(\frac{1}{4})^{r_1} (\frac{1}{4})^{r_2} (\frac{1}{4})^{r_3} 8p_i^2 p_j p_k$
7. $A_i A_j \times A_k A_l$	0

Only one parental cross can produce a set of sibs with these genotypes, so

$$\Pr(\Delta_8 \mid R_{1:n}) = \left(\frac{1}{4}\right)^{r_1+r_2+r_3} 8p_i^2 p_j p_k = 2 \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j p_k.$$

Relationship $R_{n:1,1,\dots}$

$$\Pr(\Delta_8 \mid R_{n:1,1,\dots}) = p_i^{2r_1} (2p_i p_j)^{r_2} (2p_i p_k)^{r_3} = 2^{n-r_1} p_i^{n+r_1} p_j^{r_2} p_k^{r_3}.$$

Family Pattern Δ_9 : There are r_1 individuals with genotype $A_i A_i$, r_2 with $A_i A_j$ and r_3 with $A_j A_k$ for a family pattern Δ_9 .

Relationship $R_{1:n}$

Parent Pair	$\Pr(\Delta_9 \mid \text{Parent Pair})$
1. $A_i A_i \times A_i A_i$	0
2. $A_i A_i \times A_j A_j$	0
3. $A_i A_i \times A_i A_j$	0
4. $A_i A_i \times A_j A_k$	0
5. $A_i A_j \times A_i A_j$	0
6. $A_i A_j \times A_i A_k$	$(\frac{1}{4})^{r_1} (\frac{1}{4})^{r_2} (\frac{1}{4})^{r_3} 8p_i^2 p_j p_k$
7. $A_i A_j \times A_k A_l$	0

As above, only one parental cross can produce a set of sibs with these genotypes, so

$$\Pr(\Delta_9 \mid R_{1:n}) = \left(\frac{1}{4}\right)^{r_1+r_2+r_3-1} 2p_i^2 p_j p_k = 2 \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j p_k.$$

Relationship $R_{n:1,1,\dots}$;

$$\begin{aligned} \Pr(\Delta_9 \mid R_{n:1,1,\dots}) &= \prod_{x=1}^n \Pr(\gamma_x) = p_i^{2r_1} (2p_i p_j)^{r_2} (2p_j p_k)^{r_3} \\ &= 2^{n-r_1} p_i^{2r_1+r_2} p_j^{r_2+r_3} p_k^{r_3}. \end{aligned}$$

Family Pattern Δ_{10} : A family of pattern Δ_{10} has r_1 individuals with genotype $A_i A_i$, r_2 with $A_i A_j$, r_3 with $A_i A_k$, and r_4 individuals with genotype $A_j A_k$.

Relationship $R_{1:n}$;

Parent Pair	$\Pr(\Delta_9 \mid \text{Parent Pair})$
1. $A_i A_i \times A_i A_i$	0
2. $A_i A_i \times A_j A_j$	0
3. $A_i A_i \times A_i A_j$	0
4. $A_i A_i \times A_j A_k$	0
5. $A_i A_j \times A_i A_j$	0
6. $A_i A_j \times A_i A_k$	$\left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} \left(\frac{1}{4}\right)^{r_3} \left(\frac{1}{4}\right)^{r_4} 8p_i^2 p_j p_k$
7. $A_i A_j \times A_k A_l$	0

As for the previous families, only one parental cross can produce a family with these genotypes, and

$$\Pr(\Delta_{10} \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i^2 p_j p_k.$$

Relationship $R_{n:1,1,\dots}$;

$$\begin{aligned} \Pr(\Delta_{10} \mid R_{n:1,1,\dots}) &= \prod_{x=1}^n \Pr(\gamma_x) = p_i^{2r_1} (2p_i p_j)^{r_2} (2p_i p_k)^{r_3} (2p_j p_k)^{r_4} \\ &= 2^{n-r_1} p_i^{2r_1+r_2+r_3} p_j^{r_2+r_4} p_k^{r_3+r_4}. \end{aligned}$$

Family Pattern Δ_{11} : There are r_1 individuals with genotype $A_i A_j$, r_2 with genotype $A_i A_k$, and r_3 with genotype $A_j A_k$ in a Δ_{11} sample.

Relationship $R_{1:n}$;

Parent Pair	$\Pr(\Delta_{11} \mid \text{Parent Pair})$	Example for A_1A_2 , A_1A_3 and A_2A_3
1. $A_iA_i \times A_iA_i$	0	—
2. $A_iA_i \times A_jA_j$	0	—
3. $A_iA_i \times A_iA_j$	0	—
4. $A_iA_i \times A_jA_k$	0	—
5. $A_iA_j \times A_iA_j$	0	—
6. $A_iA_j \times A_iA_k$	$\left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} \left(\frac{1}{4}\right)^{r_3} 8p_i^2 p_j p_k$	$A_1A_2 \times A_1A_3$
	$+ \left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} \left(\frac{1}{4}\right)^{r_3} 8p_i p_j^2 p_k$	$A_1A_2 \times A_2A_3$
	$+ \left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} \left(\frac{1}{4}\right)^{r_3} 8p_i p_j p_k^2$	$A_1A_3 \times A_2A_3$
7. $A_iA_j \times A_kA_l$	0	—

Summing over all possible parent pairs,

$$\Pr(\Delta_{11} \mid R_{1:n}) = \left(\frac{1}{4}\right)^{r_1+r_2+r_3} 8p_i p_j p_k (p_i + p_j + p_k) = 2 \left(\frac{1}{4}\right)^{n-1} p_i p_j p_k (p_i + p_j + p_k).$$

Relationship $R_{n:1,1,\dots}$;

$$\begin{aligned} \Pr(\Delta_{11} \mid R_{n:1,1,\dots}) &= \prod_{x=1}^n \Pr(\gamma_x) = (2p_i p_j)^{r_1} (2p_i p_k)^{r_2} (2p_j p_k)^{r_3} \\ &= 2^n p_i^{r_1+r_2} p_j^{r_1+r_3} p_k^{r_2+r_3}. \end{aligned}$$

Family Pattern Δ_{12} : A Δ_{12} sample has r_1 individuals with genotype A_iA_k , r_2 with genotype A_iA_l , and r_3 with genotype A_jA_k .

Relationship $R_{1:n}$;

Parent Pair	$\Pr(\Delta_1 \mid \text{Parent Pair})$
1. $A_iA_i \times A_iA_i$	0
2. $A_iA_i \times A_jA_j$	0
3. $A_iA_i \times A_iA_j$	0
4. $A_iA_i \times A_jA_k$	0
5. $A_iA_j \times A_iA_j$	0
6. $A_iA_j \times A_iA_k$	0
7. $A_iA_j \times A_kA_l$	$\left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} \left(\frac{1}{4}\right)^{r_3} 8p_i p_j p_k p_l$

There is only one parental cross that can produce this set of genotypes, so

$$\Pr(\Delta_{12} \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} 8p_i p_j p_k p_l.$$

Relationship $R_{n:1,1,\dots}$;

$$\Pr(\Delta_{12} \mid R_{n:1,1,\dots}) = \prod_{x=1}^n \Pr(\gamma_x) = (2p_i p_k)^{r_1} (2p_i p_l)^{r_2} (2p_j p_k)^{r_3}.$$

Family Pattern Δ_{13} : There are two types of individual in this family pattern, r_1 with genotype $A_i A_k$ and r_2 with genotype $A_j A_l$.

Relationship $R_{1:n}$;

Parent Pair	$\Pr(\Delta_1 \mid \text{Parent Pair})$
1. $A_i A_i \times A_i A_i$	0
2. $A_i A_i \times A_j A_j$	0
3. $A_i A_i \times A_i A_j$	0
4. $A_i A_i \times A_j A_k$	0
5. $A_i A_j \times A_i A_j$	0
6. $A_i A_j \times A_i A_k$	0
7. $A_i A_j \times A_k A_l$	$\left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} 8p_i p_j p_k p_l$

$$\Pr(\Delta_{13} \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i p_j p_k p_l.$$

Relationship $R_{n:1,1,\dots}$;

$$\Pr(\Delta_{13} \mid R_{n:1,1,\dots}) = \prod_{x=1}^n \Pr(\gamma_x) = (2p_i p_k)^{r_1} (2p_j p_l)^{r_2}.$$

Family Pattern Δ_{14} : This final pattern, Δ_{14} , is the only one to have four different genotypes represented, r_1 of $A_i A_k$, r_2 of $A_i A_l$, r_3 of $A_j A_k$, and r_4 of genotype $A_j A_l$.

Relationship $R_{1:n}$;

Parent Pair	$\Pr(\Delta_1 \mid \text{Parent Pair})$
1. $A_i A_i \times A_i A_i$	0
2. $A_i A_i \times A_j A_j$	0
3. $A_i A_i \times A_i A_j$	0

$$\begin{array}{ll} 4. & A_i A_i \times A_j A_k \quad 0 \\ 5. & A_i A_j \times A_i A_j \quad 0 \\ 6. & A_i A_j \times A_i A_k \quad 0 \\ 7. & A_i A_j \times A_k A_l \quad \left(\frac{1}{4}\right)^{r_1} \left(\frac{1}{4}\right)^{r_2} \left(\frac{1}{4}\right)^{r_3} \left(\frac{1}{4}\right)^{r_4} 8p_i p_j p_k p_l \end{array}$$

$$\Pr(\Delta_{14} \mid R_{1:n}) = 2 \left(\frac{1}{4}\right)^{n-1} p_i p_j p_k p_l.$$

Relationship $R_{n:1,1,\dots}$;

$$\Pr(\Delta_{14} \mid R_{n:1,1,\dots}) = \prod_{x=1}^n \Pr(\gamma_x) = (2p_i p_k)^{r_1} (2p_i p_l)^{r_2} (2p_j p_k)^{r_3} (2p_j p_l)^{r_4}.$$

Appendix B

Computer Programs

Random Sampling and Computer Simulations

Simulation is used when it is not possible, or is impractical, to derive a solution to a problem by analytical methods, to investigate the structure of realizations or to back up theory [Mize & Cox, 1968]. The genotype simulation algorithm used for both analyses in this thesis performs two main functions: to select the parent individuals and to choose the genotypes of their offspring. The sample should be representative of the population from which it is drawn, i.e. a population of individuals that are in Hardy-Weinberg equilibrium with respect to the loci of interest. The parent-pairs are chosen by random sampling from the population of parents and the offspring by random sampling of the gametes. The frequency of each type of parent-pair and their gametes depends on the gene frequency for the locus under consideration.

Given a sequence of independent observations of a random variable whose distribution is known, it is possible to construct a sequence of independent observations of any other random variable whose distribution is known. The first sequence is usually, for simplicity, a sequence of random numbers from a uniform distribution and is known as the basic sequence. When a basic sequence is used like this to generate a sequence of observations of another random variable with a known distribution, it is called *Monte Carlo* sampling.

Random sampling from a uniform or rectangular distribution over the interval 0 to 1 can be made equivalent to random sampling from any known distribution. Suppose $E_1, E_2, \dots, E_n$ are the possible outcomes from an experiment and the probability that each occurs is $p_1, p_2, \dots, p_n$ where $\sum p_i = 1$. For a given trial, a uniform pseudo-random number x_i is generated ($0 \leq x_i \leq 1$). If $x \leq p_1$ then E_1 occurs. E_2 occurs if

$p_1 < x \leq p_1 + p_2$, etc i.e.

$$E_j \text{ occurs if } \sum_{i=1}^{j-1} p_i < x \leq \sum_{i=1}^j p_i, \quad j = 2, 3, \dots, n.$$

It is assumed that successive trials are independent and that the density function of the random variable remains unchanged for successive repetitions.

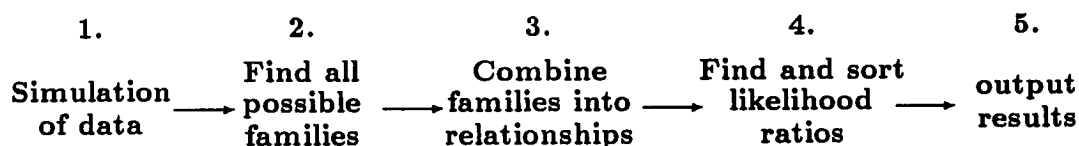
Random numbers may be selected from tables or generated by computer subroutine. It is useful to be able to repeat calculations using the same sample as a method of checking results. This means that it must be possible to repeat the same sequence of 'random' numbers. It would be impractical to store all numbers used so this is achieved by generating a sequence of pseudo-random numbers by following a calculation rule, usually using a computer program. The sequence can be repeated by using a known seed.

B.1 The Likelihood Method Programs

Throughout Chapter 3, theoretical results are verified using computer simulation, whilst in Chapter 4, conclusions are based entirely on results obtained from the same programs. The main set of programs used to generate and analyse genotype data are described in this section.

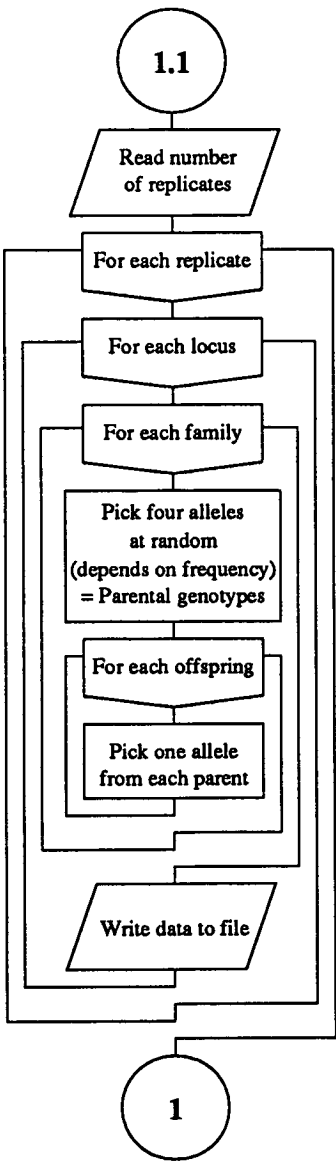
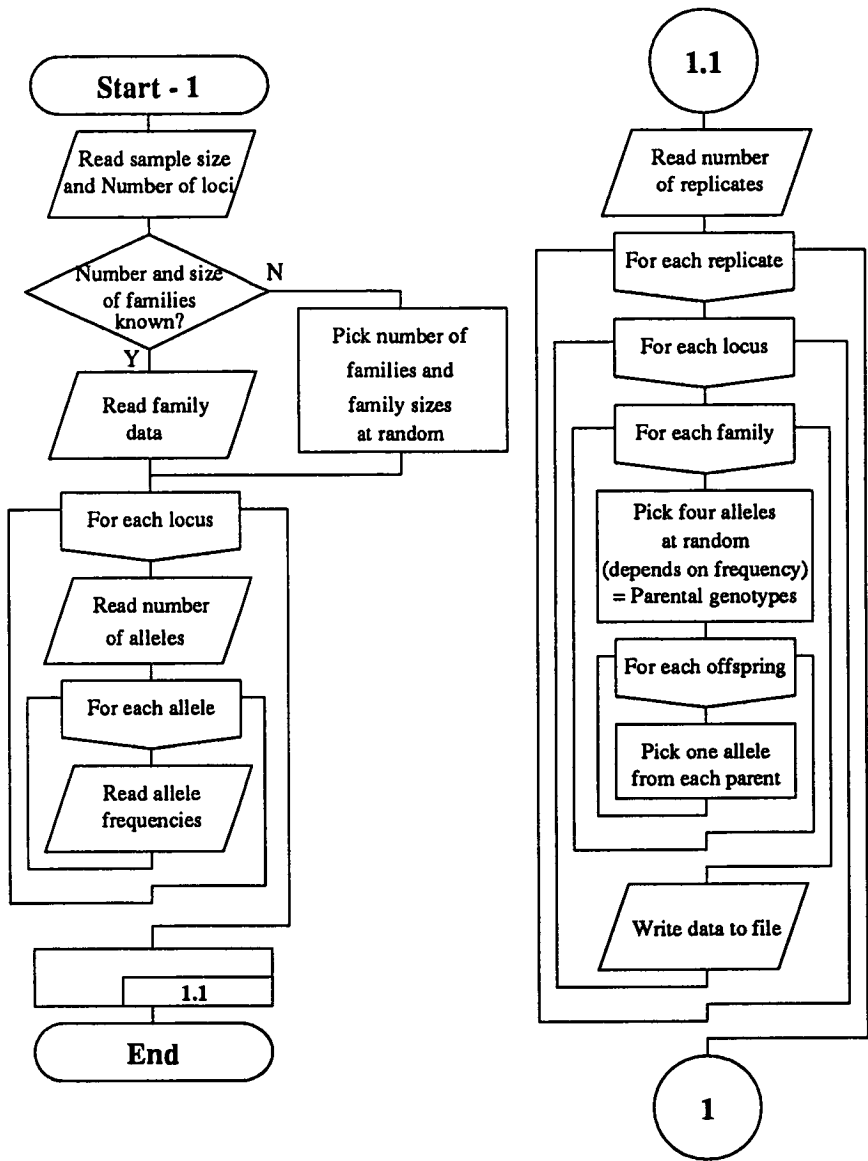
In order to keep the method flexible, a number of separate programs were written, one for each stage of the analysis. Each program can be run in isolation to carry out a particular subset of calculations; when used consecutively, each program is designed to read in data from the previous program, process the information and pass the results to the next program in the sequence. In this way, intermediate results are readily examined and each program can be modified or replaced provided the input and output formats are held constant. The use of several small programs also enables the analysis to be carried out on a PC as well as on a mainframe computer.

There are five main programs;

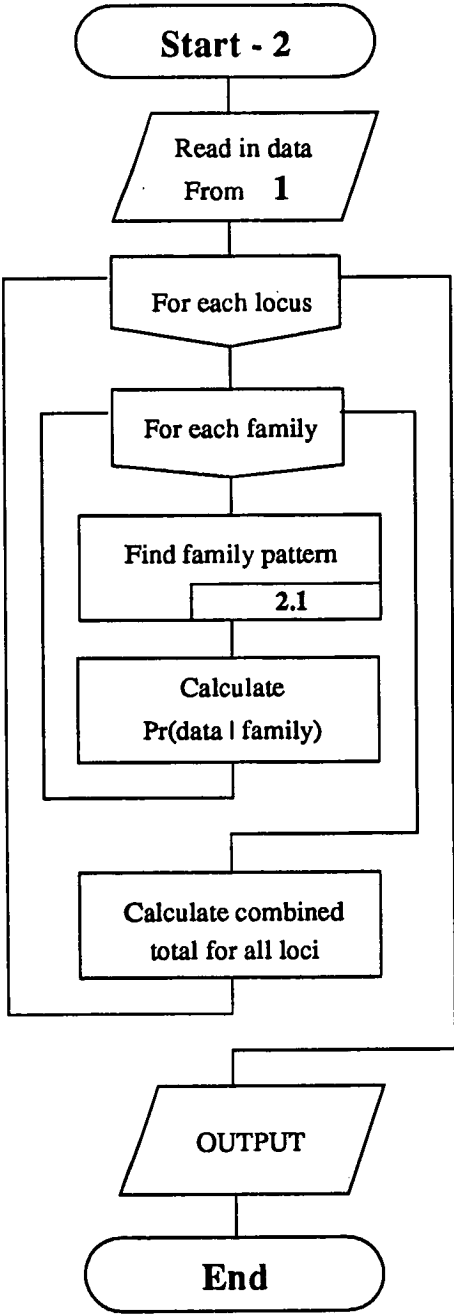


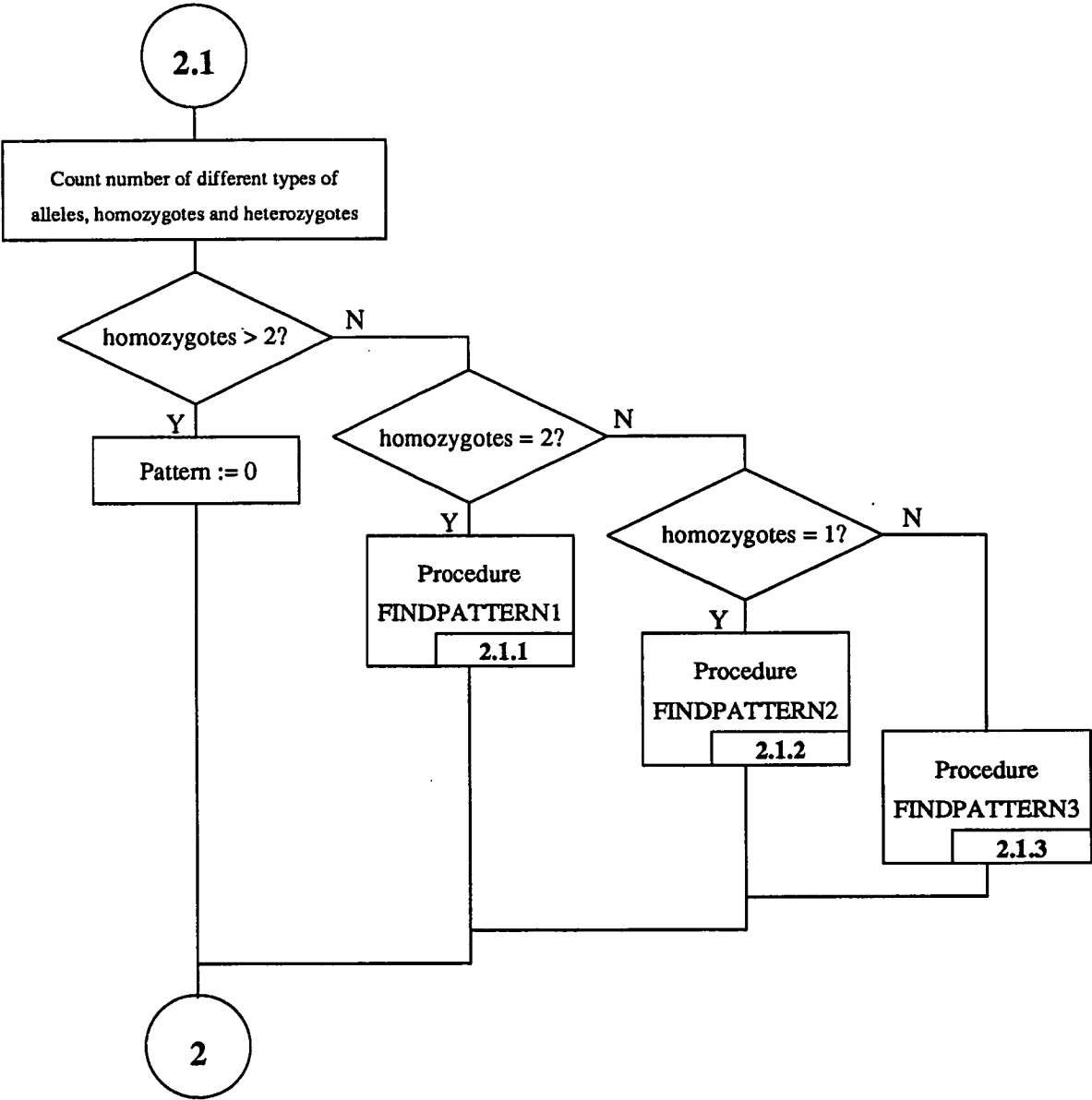
Each program is now described briefly, and a flow chart is given showing the main function of each of the programs.

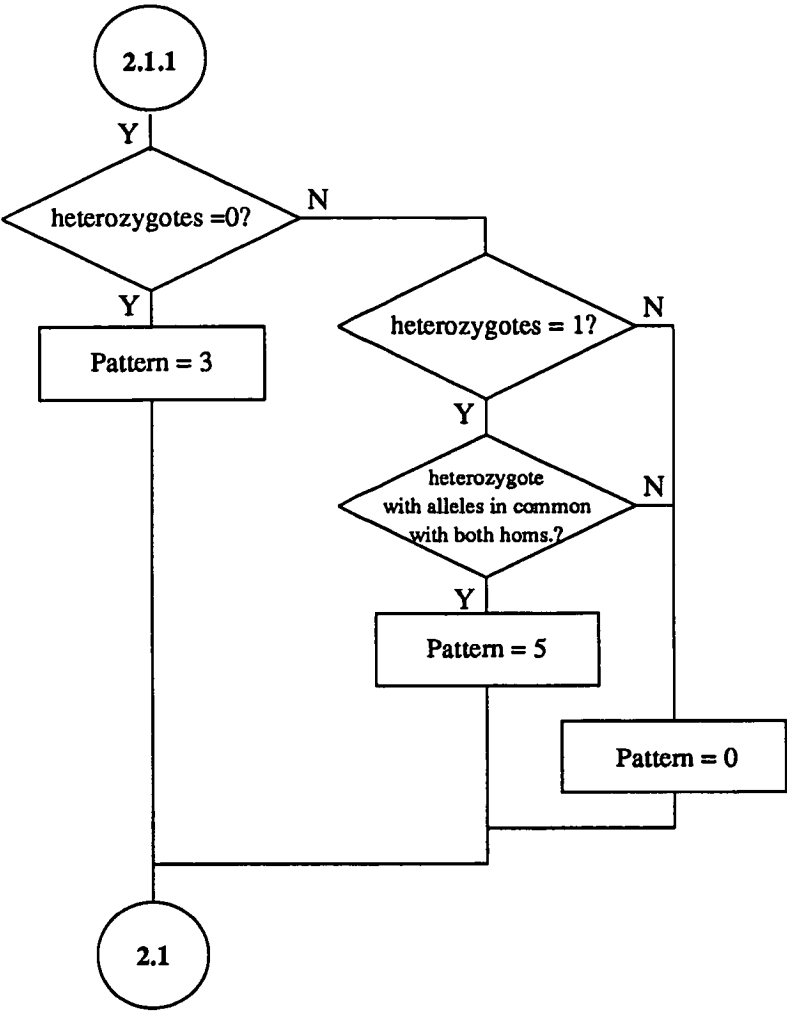
1. Program SIMUL. The function of this first program is to generate genotypes for the individuals in the sample. The first part of the program reads in the required parameters: the sample size, the number of loci and the number and frequency of alleles per locus. The required relationship structure may be given explicitly or chosen at random from the set of possible relationships for the given sample size. The second part of the program uses Monte-Carlo simulation to select the four alleles per locus that form the genotypes of the parents for each family B. Then, for each locus and each family, pairs of alleles are selected at random from these four “parental” alleles to form the genotypes of the required number of offspring.

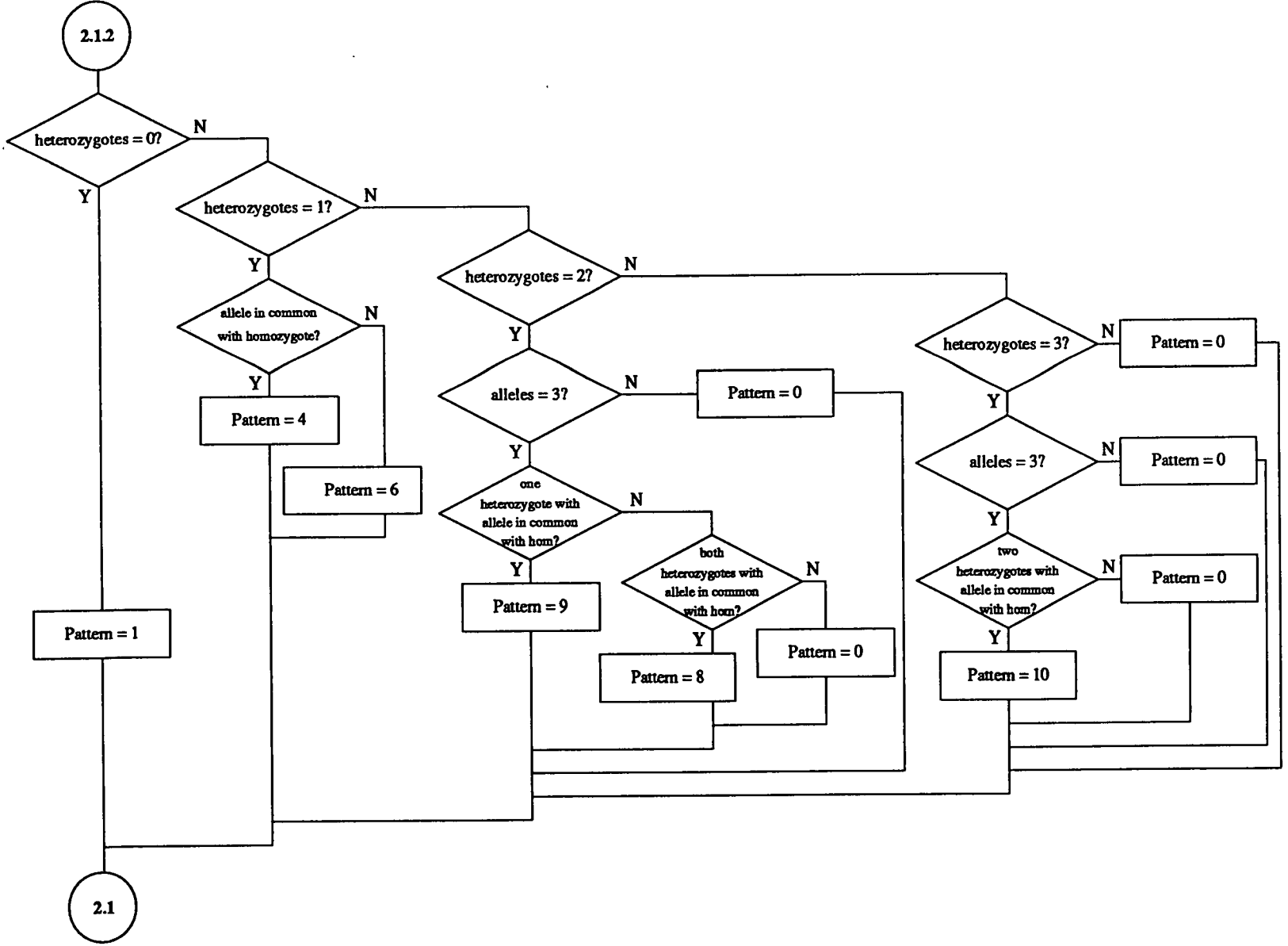


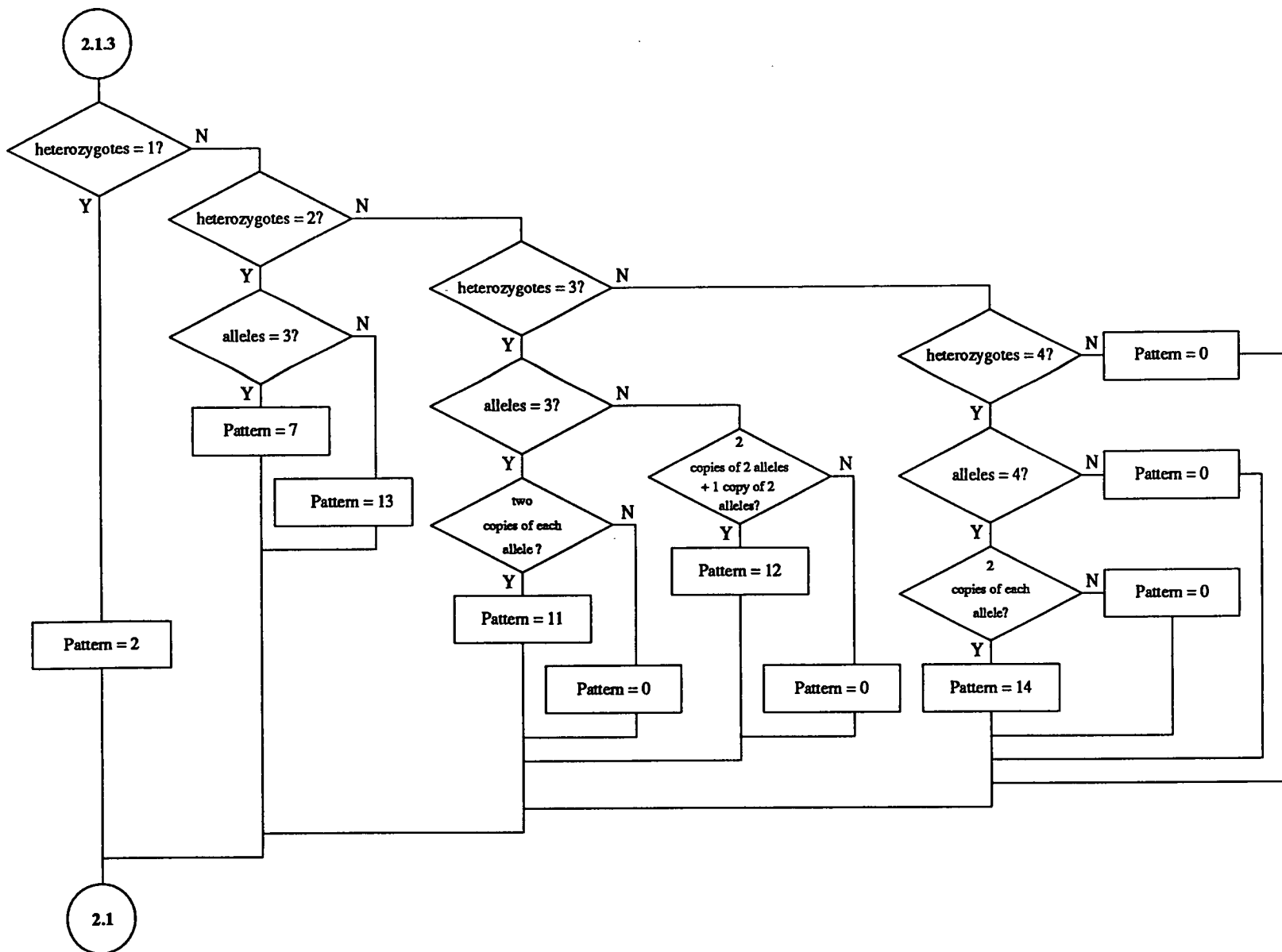
2. Program FAMILY. The second program in the sequence reads in the genotypes (either from program SIMUL or interactively) and finds the probability of obtaining these observed genotypes for all possible combinations of sibs within families. That is, the set of Expressions (2.5) to (2.31) are evaluated for valid families.



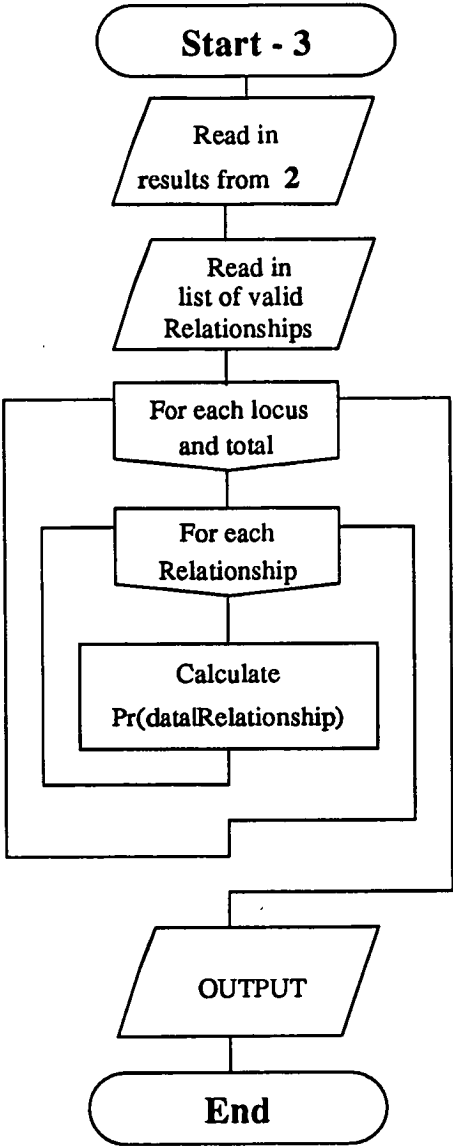




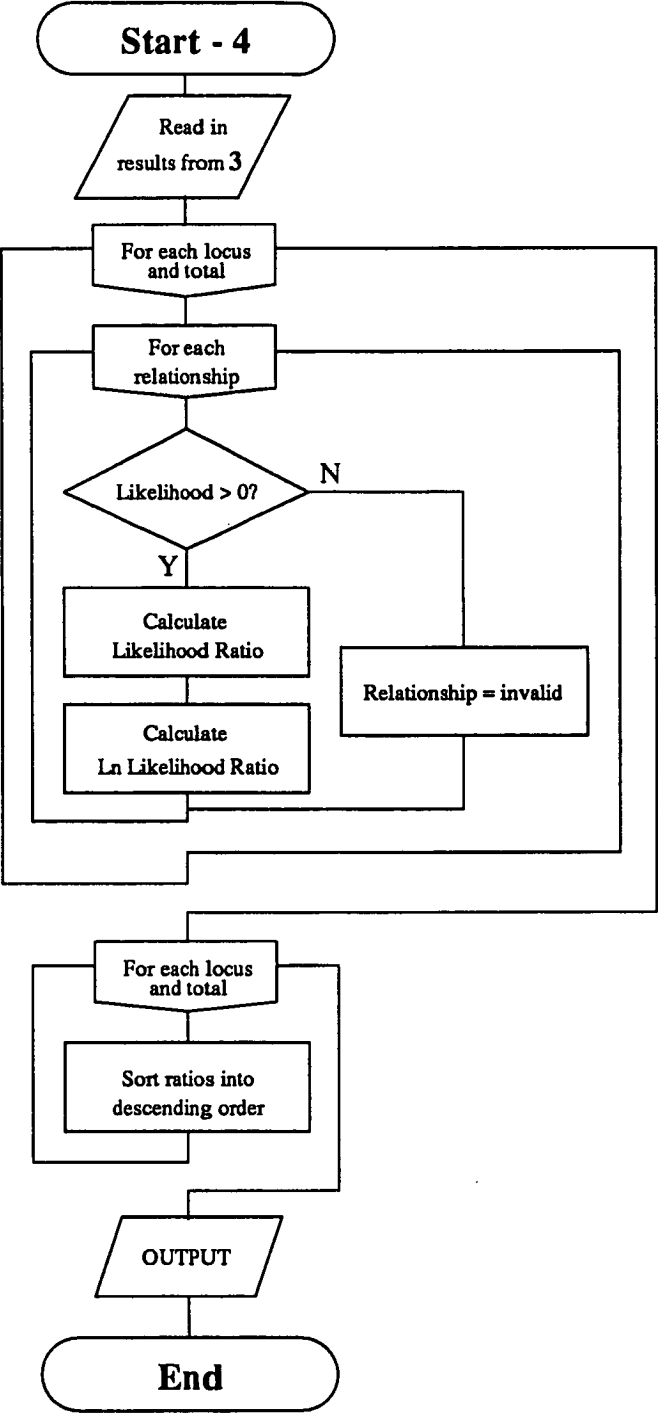




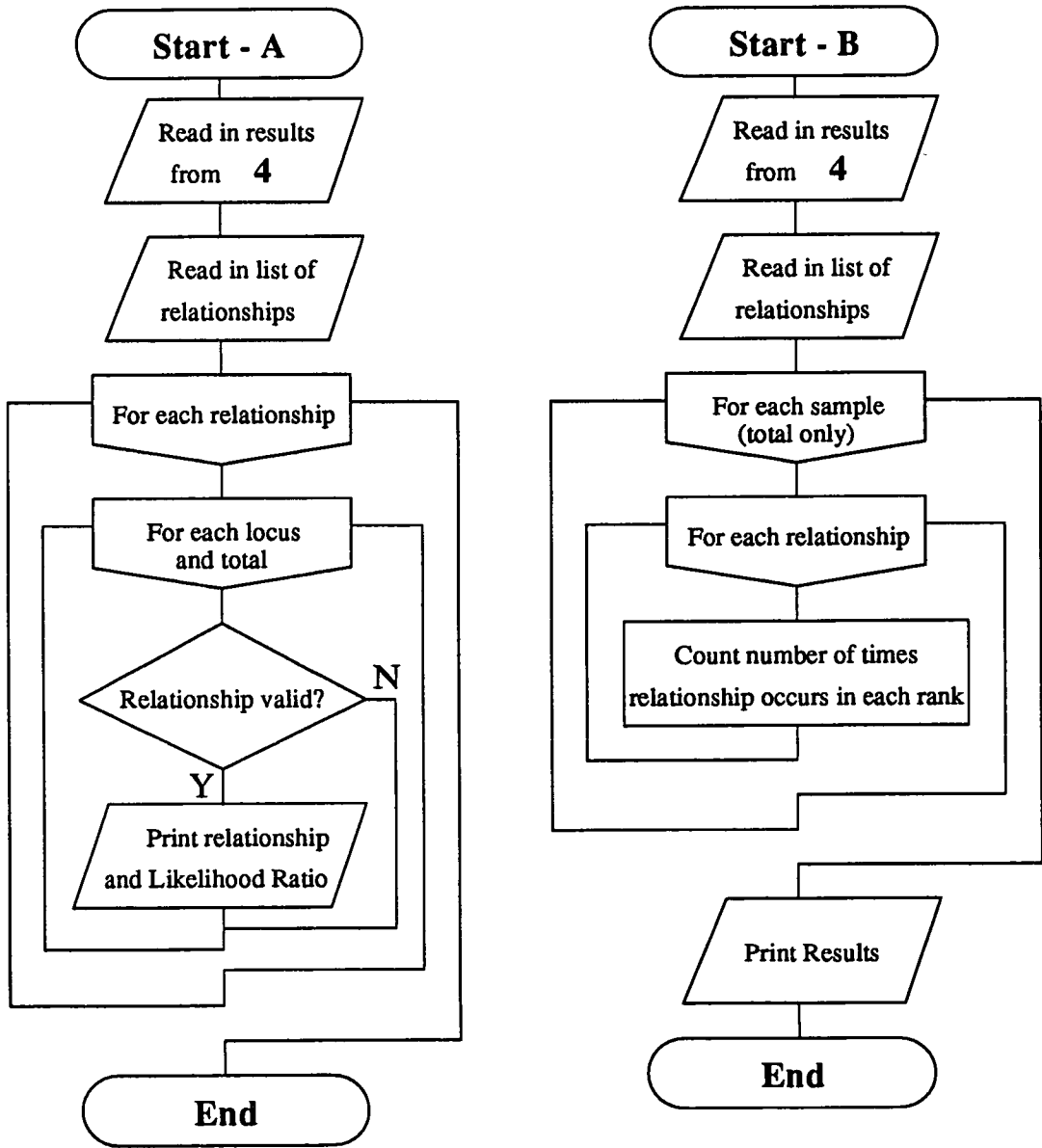
3. Program LIKELIHOOD. This short program takes the probabilities for each set of sibs and combines the appropriate families into valid relationships as described in Section 2.3.4.



4. **Program RANK.** This program reads in the likelihoods for each relationship and calculates the likelihood ratio (relative to that for $R_{N:1,1,\dots}$) and the natural logarithm of the likelihood ratio. These ratios are then sorted into descending order of magnitude, each relationship being assigned a rank (see Section 2.4).

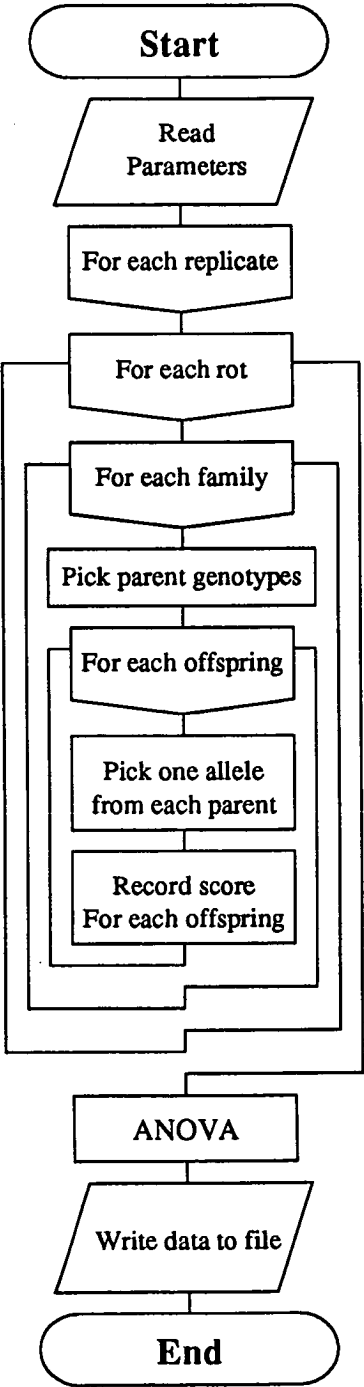


5. Output Programs. The exact format of the output results can vary depending on the type of analysis, so there are a number of different programs that can read in the output from Program RANK and process the results accordingly. Two such programs are shown below. Program A simply prints out the ranked likelihoods for each locus and the combined total together with a description of each relationship structure. This type of output is displayed in Section 5.2. Program B is used to combine results from many replicate samples and calculate the probability of correct inference as described in Section 3.2.2.



B.2 Analysis of Variance Program

In Chapter 6, a number of simulation experiments were carried out. The flowchart below outlines the procedure followed in the program ANOSIM. The genotypes for the required number of individuals are generated and the scores found depending on the indicator variable used (based on alleles, genotypes or heterozygotes). An analysis of variance is carried out, and results recorded in a file.



Bibliography

- Abramowitz, M. and Stegun, I.A. (1965). *Handbook of Mathematical Functions*. AMS 55 U.S. Department of Commerce, Dover edition.
- Baird, E., Cooper-Bland, S., Waugh, R., DeMaine, M. and Powell, W. (1992). Molecular characterisation of inter- and intra-specific somatic hybrids of potato using randomly amplified polymorphic DNA (RAPD) markers. *Molecular and General Genetics*, 233:469–475.
- Barker, J.S.F. (1992). Genetic variation in cactophilic *Drosophila* for oviposition on natural yeast substrates. *Evolution*, 46:1070–1083.
- Barker, J.S.F., East, P.D. and Weir, B.S. (1986). Temporal and microgeographic variation in allozyme frequencies in a natural population of *Drosophila buzzatii*. *Genetics*, 112:577–611.
- Budowle, B., Giusti, A.M., Wayne, J.S., Baechtel, F.S., Fournay, R.M., Adams, D.E., Presley, L.A., Deadman, H.A. and Monson, K.L. (1991). Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. *American Journal of Human Genetics*, 48:841–855.
- Burke, T., Davies, N.B., Bruford, M.W. and Hatchwell, B.J. (1989). Parental care and mating behaviour of polyandrous dunnocks *Prunella modularis* related to paternity by DNA fingerprinting. *Nature*, 338:249–251.
- Burke, T., Hanotte, O., Bruford, M.W. and Cairns, E. (1991). Multilocus and single locus minisatellite analysis in population biological studies. In Burke, T., Dolf, G., Jeffreys, A.J. and Wolff, R., (eds.), *DNA Fingerprinting: Approaches and Applications*, pages 154–168. Birkhäuser.
- Cockerham, C.C. (1969). Variance of gene frequencies. *Evolution*, 23:72–84.
- Debenham, P.G. (1992). Probing identity: the changing face of DNA fingerprinting. *Trends in Biotechnology*, 10:96–102.

- Devlin, B., Risch, N. and Roeder, K. (1991). Estimation of allele frequencies for VNTR loci. *American Journal of Human Genetics*, 48:662-676.
- Edwards, A.W.F. (1972). *Likelihood: An account of the statistical concept of likelihood*. Cambridge University Press.
- Elandt-Johnson, R.C. (1971a). Joint genotype distributions of s children and a parent and of s siblings. Multiple alleles. *American Journal of Human Genetics*, 28:442-461.
- Elandt-Johnson, R.C. (1971b). *Probability Models and Statistical methods in genetics*. Wiley.
- Elston, R.C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity*, 21:523-542.
- Fisher, R.A. (1941). *Statistical Methods for Research Workers*. Oliver & Boyd, 8th edition.
- Foltz, D.W. and Hoogland, J.L. (1981). (no title). *Journal of Mammalogy*, 63:706-712.
- Gill, P., Jeffreys, A.J. and Werrett, D.J. (1985). Forensic application of DNA fingerprints. *Nature*, 318:577-579.
- Hagelberg, E., Gray, I.C. and Jeffreys, A.J. (1991). Identification of the skeletal remains of a murder victim by DNA analysis. *Nature*, 352:427-429.
- Hanken, J. and Sherman, P.W. (1981). Multiple paternity in Belding's ground squirrel litters. *Science*, 212:351-351.
- Hearn, A.C. (1987). *REDUCE Users Manual*. Rand Corporation, 3.3 edition.
- Hill, W.G. (1987). DNA fingerprints applied to animal and bird populations. *Nature*, 327:98-99.
- Hoffmann, A.A. and Nielsen, K.M. (1985). The effect of resource subdivision on genetic variation in *Drosophila*. *American Naturalist*, 125:421-430.
- Hunt, G.J. and Page, R.E.Jr. (1992). Patterns of inheritance with RAPD molecular markers reveal novel types of polymorphism in the honey bee. *Theoretical and Applied Genetics*, 85:15-20.
- Jacquard, A. (1972). Genetic information given by a relative. *Biometrics*, 28:1101-1114.

- Jaenike, J. and Selander, R.K. (1979). Enforced generalism in mycophagous *Drosophila*: genetic evidence. *Evolution*, 33:741-749.
- Jeffreys, A.J. and Morton, D.B. (1987). DNA fingerprints of dogs and cats. *Animal Genetics*, 18:1-15.
- Jeffreys, A.J., Brookfield, J.F.Y. and Semeonoff, R. (1985a). Positive identification of an immigration test-case using human DNA fingerprints. *Nature*, 317:818-819.
- Jeffreys, A.J., Wilson, V. and Thein, S.L. (1985b). Hypervariable "minisatellite" regions in human DNA. *Nature*, 314:67-73.
- Jeffreys, A.J., MacLeod, A., Tamaki, K., Neil, D.L. and Monckton, D.G. (1991a). Minisatellite repeat coding as a digital approach to DNA typing. *Nature*, 354:204-209.
- Jeffreys, A.J., Royle, N.J., Patel, I., Armour, J.A.L., MacLeod, A., Collick, A., Gray, I.C., Neumann, R., Gibbs, M., Crosier, M., Hill, M., Signer, E. and Monckton, D. (1991b). Principles and recent advances in human DNA fingerprinting. In Burke, T., Dolf, G., Jeffreys, A.J. and Wolff, R., (eds.), *DNA Fingerprinting: Approaches and Applications*, pages 1-19. Birkhäuser.
- Jeffreys, A.J., Turner, M. and Debenham, P. (1991c). The efficiency of multilocus DNA fingerprint probes for individualisation and establishment of family relationships, determined from extensive casework. *American Journal of Human Genetics*, 48:824-840.
- Kuhnlein, U., Zadworny, D., Dawe, Y., Fairfull, R.W. and Gavora, J.S. (1990). Assessment of inbreeding by DNA fingerprinting: Development of a calibration curve using defined strains of chickens. *Genetics*, 125:161-165.
- Lacy, R.C. (1983). Structure of genetic variation within and between populations of mycophagous *Drosophila*. *Genetics*, 104:81-94.
- Lynch, M. (1988). Estimation of relatedness by DNA fingerprinting. *Molecular Biology and Evolution*, 5(5):584-599.
- Lynch, M. (1991). Analysis of population genetic structure by DNA fingerprinting. In Burke, T., Dolf, G., Jeffreys, A.J. and Wolff, R., (eds.), *DNA Fingerprinting: Approaches and Applications*, pages 113-126. Birkhäuser.

- Meagher, T.R. and Thompson, E.A. (1986). The relationship between single parent and parent pair genetic likelihoods in genealogical reconstruction. *Theoretical Population Biology*, 29:87-106.
- Meagher, T.R. and Thompson, E.A. (1987). Analysis of parentage for naturally established seedlings of *Chamaelirium luteum*. *Ecology*, 68(4):803-812.
- Mize, J.H. and Cox, J.G. (1968). *Essentials of Simulation*. Prentice-Hall.
- Morton, D.B., Yaxley, R.E., Patel, I., Jeffreys, A.J., Howes, S.J. and Debenham, P.G. (1987). (no title). *Veterinary Record*, 121:592-593.
- Pellissier Scott, M., Haymes, K.M. and Williams, S.M. (1992). Parentage analysis using RAPD PCR. *Nucleic Acid Research*, 20:5493.
- Quinn, T.W., Quinn, J.S., Cooke, F. and White, B.N. (1987). DNA marker analysis detects multiple maternity and paternity in single broods of the lesser snow goose. *Nature*, 326:392-394.
- Roeder, K., Devlin, B. and Lindsay, B.G. (1989). Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities. *Biometrics*, 45:363-379.
- Schwarzmann, J.F. and Gerhold, H.D. (1991). Genetic structure and mating system of northern red oak (*Quercus rubra* L) in Pennsylvania. *Forest Science*, 37:1376-1389.
- Thomas, R.H. and Barker, J.S.F. (1990). Breeding structure of natural populations of *Drosophila buzzatii*: effects of the distribution of larval substrates. *Heredity*, 64:355-365.
- Thompson, E.A. (1974a). Gene identities and multiple relationships. *Biometrics*, 30:667-680.
- Thompson, E.A. (1974b). *Mathematical Analysis of human evolution and population structure*. Ph.D. thesis (unpublished), University of Cambridge, England.
- Thompson, E.A. (1975). The estimation of pairwise relationships. *Annals of Human Genetics*, 39:173-187.
- Thompson, E.A. (1976a). Inference of genealogical structure. *Social Science Information*, 15:477-526.

- Thompson, E.A. (1976b). A paradox of genealogical inference. *Advanced Applied Probability*, 8:648–650.
- Thompson, E.A. (1980). The gene identity states of a descendant. *Theoretical Population Biology*, 18:76–93.
- Thompson, E.A. (1986). Likelihood inference of paternity. *American Journal of Human Genetics*, 39:285–287.
- Thompson, E.A., Cannings, C. and Skolnick, M.H. (1978). Ancestral inference: I. The problem and the method. *Annals of Human Genetics*, 42:95–108.
- VandeBerg, J.L., Aivaliotis, M.J., Williams, L.E. and Abee, C.R. (1990). Biochemical genetic markers of squirrel monkeys and their use for pedigree validation. *Biochemical Genetics*, 28:41–56.
- Vierling, R.A. and Nguyen, H.T. (1992). Use of RAPD markers to determine the genetic diversity of diploid wheat genotypes. *Theoretical and Applied Genetics*, 84:835–838.
- Weber, J.L. and May, P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using polymerase chain reaction. *American Journal of Human Genetics*, 44:388–396.
- Weir, B.S. and Cockerham, C.C. (1984). Estimating f -statistics for the analysis of population structure. *Evolution*, 38:1358–1370.
- Weir, B.S. (1992). Independence of VNTR alleles defined as fixed bins. *Genetics*, 130:873–887.
- Wells, R.A., Wonke, B. and Thein, S.L. (1988). Prediction of consanguinity using human DNA fingerprints. *Journal of Medical Genetics*, 25:660–662.
- Welsh, J. and McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acid Research*, 18:7213–7218.
- Wetton, J.H., Carter, R.E., Parkin, D.T. and Walters, D. (1987). Demographic study of a wild house sparrow population by DNA fingerprinting. *Nature*, 327:147–149.
- Williams, C.J. and Evarts, S. (1989). The estimation of concurrent multiple paternity probabilities in natural populations. *Theoretical Population Biology*, 35:90–112.

- Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A. and Tingey, S.V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acid Research*, 18:6531-6535.
- Wong, A., Wilson, V., Patel, I., Povey, S. and Jeffreys, A.J. (1987). Characterisation of highly variable minisatellites cloned from human DNA. *Annals of Human Genetics*, 51:269-288.
- Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, 15:323-354.