Chapter 3

Gene Identity by Descent

3.1 Kinship and inbreeding coefficients

A *gene*, as opposed to an allele or a locus, is the *DNA segment* that is copied from parents to offspring. Underlying the patterns of phenotypes observed on related individuals are the *genotypes*, but underlying the genotypes are the patterns of gene identity by descent. Phenotypes of relatives are similar because they have similar genotypes and may share a common environment. Genotypes are similar because relatives share genes that are identical by descent (*ibd*) — identical copies of a gene segregating from a common ancestor within the defined pedigree. Although for some microsatellite DNA markers mutation rates are non-negligible (section 1.1), for simplicity we disregard mutation throughout this book. In this case, genes that are *ibd* must be of the same allelic type, while genes that are not *ibd* are of independent allelic types.

Gene identity by descent is defined only within the context of a given pedigree structure. A pedigree specifies the two parents of every non-founder individual. A founder has neither parent specified, and by definition the genes in founders are not *ibd*. It will often be convenient if a pedigree is ordered in such a way that every individual is preceded in the listing by his parents; clearly, this is always possible.

Mendel's First Law (section 1.2) states that:

> a diploid individual receives at any given locus a copy of a randomly chosen one of the two genes in his father and (independently) a copy of a randomly chosen one of the two genes in his mother, and will pass on a copy of a randomly and independently chosen one of these two genes to each of his offspring.

This simple law leads to complex patterns of gene identity on an extended pedigree, due to the huge number of alternative events; 2^m for m meioses, at each locus. The segregating genes determine the patterns of gene identity by descent on the pedigree, and hence the patterns of similarity among relatives.

We start with coefficients of *inbreeding* and *kinship*, since these provide an introduction to the ideas of gene identity by descent, to alternative computational

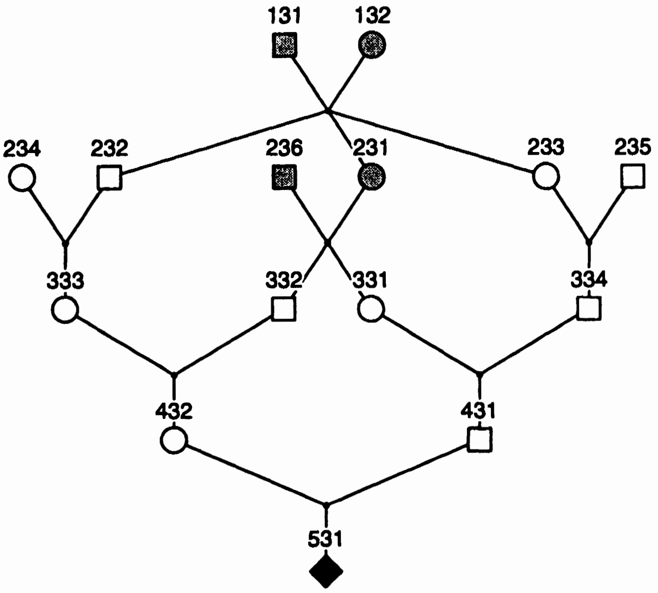

FIGURE 3.1. *An example pedigree. The structure is the same as that of Figure 1.1 of section 1.3. The four individuals shaded grey are bilateral ancestors of the final individual*

approaches, and to Monte Carlo estimation of expectations. Kinship and inbreeding are best thought of as relationships between gametes rather than between individuals. The coefficient of kinship between two individuals B and C, $\psi(B,C)$, is the probability that homologous genes on gametes segregating from B and from C are ***ibd***, while the inbreeding coefficient of an individual B, f_B, is the probability that homologous genes on the two gametes uniting to form individual B are ***ibd***. Hence

$$f_B = \psi(M_B, F_B)$$

where M_B and F_B are the parents of B. An individual is inbred if his parents are related. He is ***autozygous*** at a given locus if, at that locus, his two genes are ***ibd***. His inbreeding coefficient is the ***prior*** probability of this event: that is, the probability based only on the pedigree structure.

3.2 Methods of computation

There are several methods for computing kinship and inbreeding coefficients. The early approach of ***path-counting*** (Wright, 1922) simply enumerates all the possibilities in an efficient way. In order for the two genes within an individual

B to be *ibd*, they must descend from a common ancestor A of his parents. The probability that genes segregating from A in two distinct meioses are *ibd* is 1 if A has two *ibd* genes and 1/2 otherwise, or overall $f_A.1 + (1 - f_A).(1/2) \ = \ (1/2)(1 + f_A)$. If these two genes from A to two distinct offspring are *ibd*, then the probability the same genes descend to B gains a factor of 1/2 at each successive meiosis. A *path*, $\mathcal{P}_A$, is defined as a sequence of individuals from B ascending to a common ancestor A of his two parents, and descending to B again via a disjoint sequence of individuals. Each such path contributes a term $2^{-(m_M+m_F+1)}(1 + f_A)$ to the inbreeding coefficient f_B, where m_M and m_F are the number of meioses in the path from A to B's mother M and father F respectively. (One may count the two meioses from M and F to B, or the two meioses from A to his two offspring, but not both.) Now, at a single locus, the genes of B can be *ibd* via at most one such path; the paths provide a set of mutually exclusive and exhaustive events leading to B having two *ibd* genes. Thus the inbreeding coefficient of B is

$$f_B \ = \ \sum_A \sum_{\mathcal{P}_A} 2^{-(m_M(\mathcal{P}_A)+m_F(\mathcal{P}_A)+1)}(1 + f_A). \tag{3.1}$$

For example, for the offspring of a first cousin marriage, there are 2 paths, one via each of the two grandparents shared by his parents, each having $m_M = m_F = 2$, providing an inbreeding coefficient of $2 \times 2^{-5} \ = \ 1/16$. As a more complex example, consider again the pedigree of Figure 1.1 in section 1.3. The pedigree is shown again in Figure 3.1, with the common ancestors of the parents of the final individual shaded grey. The final individual is the offspring of a first cousin marriage, but so also is each of his parents. Here there are two paths via his great-grandparents, each having $m_M = m_F = 2$ as for the simple cousin marriage, and 3 paths via each of his parents' two shared great-grandparents, each with $m_M = m_F = 3$, providing a total inbreeding coefficient of $2 \times 2^{-5} \ + \ 2 \times 3 \times 2^{-7} \ = \ 7/64$.

Although the path-counting method is the simplest for small pedigrees, it becomes impractical on very large and complex pedigrees. For example, in a segment of a Hutterite pedigree considered by Thompson and Morgan (1989), there are over 1000 ancestral paths connecting the two parents of one individual. Other approaches to computation of inbreeding and kinship follow from equations based on the properties of Mendelian segregation. We use the meiosis indicators introduced in section 1.2 and consider the kinship coefficient $\psi(B, C)$ between two individuals B and C. Provided B is not an ancestor of C, we may condition on the segregation S from B, where

$$\Pr(S = 0) \ = \ \Pr(S = 1) \ = \ \frac{1}{2}.$$

If $S = 0$, the segregating gene is B's maternal gene; that is, a gene from the mother of B. If $S = 1$, the gene is B's paternal gene. Thus we obtain immediately

$$\begin{aligned} \psi(B, C) \ &= \ \psi(M_B, C)P(S = 0) \ + \ \psi(F_B, C)P(S = 1) \\ &= \ (\psi(M_B, C) \ + \ \psi(F_B, C))/2 \end{aligned} \tag{3.2}$$

where M_B and F_B are the mother and the father of B. Also, from the definition, we have symmetry: $\psi(B, C) = \psi(C, B)$. Thus the only additional equation needed is for the case $B = C$. In this case, we must consider two independent segregations from B, S_1 and S_2:

$$\Pr(S_1 = S_2) \;=\; \Pr(S_1 \neq S_2) \;=\; \frac{1}{2}.$$

If $S_1 = S_2$, the segregating genes are *ibd.* If $S_1 \neq S_2$, the genes comprise both the maternal and paternal genes of B. Thus

$$\begin{aligned}\psi(B,B) &= P(S_1 = S_2) \;+\; \psi(M_B, F_B)P(S_1 \neq S_2)\\ &= (1 \;+\; \psi(M_B, F_B))/2.\end{aligned}$$

Together with the boundary conditions

$$\begin{array}{rlll} & \psi(B,B) & = \frac{1}{2} & \text{for any founder } B,\\ \text{and} & \psi(B,C) & = 0 & \text{if } B \text{ is a founder not an ancestor of } C,\end{array}$$

these equations determine the function $\psi(\cdot)$ on the pedigree.

A recursive algorithm based on these equations is very easily implemented, and works well even on large and complex pedigrees. However, it is not necessarily computationally efficient; the same expansion may be repeated many times. In principle, this can be avoided, by saving $\psi(B,\ C)$, for key pairs of individuals (B, C) in the ancestry of the pedigree, but the simplicity of the method is then lost. An alternative way to implement these equations is via a top-down sequential method, computing kinship coefficients between all pairs of ancestors arriving finally at the descendant individuals of interest. This is computationally trivial, but expensive on store. All computation is a trade-off between time and store.

3.3 Data on inbred individuals

Kinship and inbreeding coefficients measure only *ibd* between two gametes, at a single locus. However, this suffices for a consideration of data on unrelated inbred individuals. At a single locus, with alleles $A_1, \ldots, A_k$, having population frequencies $q_1, \ldots, q_k$, an individual having two *ibd* genes has genotype A_jA_j with probability q_j, while an individual who is not autozygous at this locus has genotype probabilities of Hardy-Weinberg form (section 2.3). Thus an individual who has inbreeding coefficient f has genotype probabilities

$$\begin{aligned}\Pr(A_jA_j) &= q_jf \;+\; q_j^2(1-f)\\ &= q_j(q_j \;+\; f(1-q_j)),\;\; j = 1, \ldots, k\\ \Pr(A_jA_l) &= 2(1-f)q_jq_l,\;\; 1 \leq j < l \leq k.\end{aligned} \tag{3.3}$$

Since an individual who is autozygous at a particular locus must be homozygous at that locus, inbreeding is of particular interest in the study of rare recessive traits. If the recessive allele has frequency q, the probability that an individual with inbreeding coefficient f is affected is $q(q + f(1-q))$. If the population consists of a proportion α_i of individuals with inbreeding coefficient f_i, then the overall proportion of affected individuals is

$$\sum_i \alpha_i(q(q + f_i(1-q))) \quad = \quad q(q + f(1-q))$$

where $f \; = \; \sum_i \alpha_i f_i$ is the mean inbreeding coefficient in the population, or the expected inbreeding coefficient of an individual randomly chosen from the population. The conditional probability that an affected individual derives from the group with inbreeding coefficient f_i is

$$\frac{\alpha_i(q + f_i(1-q))}{q + f(1-q)}.$$

The probability that an affected individual with inbreeding coefficient f_i is autozygous at this locus is

$$\frac{f_i}{q + f_i(1-q)}$$

while the overall probability an affected individual is autozygous at this locus is

$$\frac{f}{q + f(1-q)}. \tag{3.4}$$

Note that for a very rare recessive trait ($q \approx 0$), a high proportion of the affected individuals will have non-zero inbreeding coefficients. Indeed, the groups i then contribute to the affected individuals in the same proportions $\alpha_i f_i / f$ as they contribute to the mean population inbreeding. Moreover, a high proportion of the affected individuals are not only inbred, but in fact autozygous at the locus in question. We return to these probabilities in section 4.6.

In a population in which the mean inbreeding coefficient is f, the genotype frequencies are given by equation (3.3). There are two points to note about this homozygote excess and heterozygote deficiency, relative to Hardy-Weinberg proportions. The first is that these are frequencies in an infinite population. In a finite population, individuals of necessity marry their relatives, and allele frequencies change over time. Whether or not there is a homozygote excess, relative to Hardy-Weinberg proportions with the current allele frequencies, depends on whether individuals are, on average, marrying an individual who is more or less closely related to them than is a randomly chosen member of the population. Second, the homozygote excess due to inbreeding is a particular special case of the homozygote excess due to subdivision of a population; inbreeding is a form of subdivision. However, under the inbreeding scenario, there is no differentiation among alleles. Under subdivision, different alleles may show differing patterns of variation in frequency among subdivisions. This leads to genotype frequencies in which each homozygote shows an excess frequency, but in an amount dependent on the variation of the frequency of that allele among subdivisions. Although in total there is a heterozygote deficiency, patterns of covariation of allele frequency may lead to increased frequencies of some heterozygote genotypes (Weir, 1996).

As an additional example of the use of the EM algorithm (section 2.4) to estimate parameters underlying genotype frequencies, we consider estimation of f under the model of equation (3.3). Suppose that a random sample of individuals is taken from the population, and that there are n_{jl} individuals of genotype A_jA_l for $j \leq l$. Then the likelihood for the parameters $\mathbf{q} = (q_1, \ldots, q_k)$ and f is

$$\begin{aligned} L(\mathbf{q}, f) &= P_{\mathbf{q},f}(\{n_{jl}\}) \\ &\propto \prod_j (q_j(q_j + f(1-q_j)))^{n_{jj}} \prod_{j<l} (2q_jq_l(1-f))^{n_{jl}}. \end{aligned}$$

Clearly, this is not an easy expression to maximize.

Let X_j be the number of homozygous A_jA_j individuals in the sample who have two identical-by-descent (*ibd*) genes at this locus. With X_j as the latent variables, the complete-data likelihood is

$$\begin{aligned} L^*(\mathbf{q}, f) &= P_{\mathbf{q},f}(\{n_{jl}\}, \{X_j\}) \\ &= \prod_j q_j^{2n_{jj}-X_j} \prod_{j<l} (2q_jq_l)^{n_{jl}} f^T (1-f)^{n-T} \end{aligned}$$

where $T = \sum_j X_j$. Let $m_j = 2n_{jj} + \sum_{l<j} n_{lj} + \sum_{l>j} n_{jl}$ be the number of A_j alleles observed in the sample. Then the complete-data log-likelihood reduces to

$$\begin{aligned} \ell^*(\mathbf{q}, f) &= \log P_{\mathbf{q},f}(\{n_{jl}\}, \{X_j\})) \\ &= \text{const} + \sum_j (m_j - X_j) \log q_j + T \log f + (n-T) \log(1-f). \end{aligned} \tag{3.5}$$

The complete-data log-likelihood (3.5) is thus linear in the functions of the latent variables X_j and T. Computation of the expected complete-data log-likelihood requires only

$$\mathrm{E}_{\mathbf{q},f}(X_j \mid \{n_{jl}\}) = \frac{f n_{jj}}{f + q_j(1-f)}$$

using equation (3.4). Moreover, if X_j were observed, the MLEs based on (3.5) would be $\widehat{f} = T/n$ and $\widehat{q}_j = (m_j - x_j)/\sum_l (m_l - x_l)$. An EM algorithm for this problem is thus to iterate:

$$\begin{aligned} \mathrm{E-step}: \quad x_j &= f n_{jj}/(f + q_j(1-f)), \quad & t &= \textstyle\sum_j x_j \\ \mathrm{M-step}: \quad q_j &= (m_j - x_j)/\textstyle\sum_l (m_l - x_l), \quad & f &= t/n. \end{aligned}$$

As in the examples of section 2.5, the algorithm is easily implemented, and converges quickly.

3.4 Multi-gamete kinship and gene *ibd*

Kinship and inbreeding provide results only concerning a pair of genes, and thus a single genotype. Analysis even of data on a pair of related individuals, at a single locus, requires consideration of four genes. An important extension to section 3.2 was made by Karigl (1981), who considered the probability of simultaneous identity by descent, $\psi(B_1, ..., B_m)$, of m genes segregating from a set of (not necessarily distinct) individuals $B_1, B_2, ..., B_m$. As in equation (3.2), if B_1 is not an ancestor of any of $B_2, ..., B_m$, conditioning on the segregation from B_1 gives

$$(3.6)\quad \psi(B_1, B_2, ..., B_m) \;=\; \frac{1}{2}\Big(\psi(M_{B_1}, B_2, ..., B_m) \;+\; \psi(F_{B_1}, B_2, ..., B_m)\Big)$$

where M_{B_1} and F_{B_1} are the parents of individual B_1. The symmetry of the definition provides that we may collect the arguments for some B_1 who is not an ancestor of any of the others to the first v arguments of ψ. Then, considering the v independent segregations from B_1, either the segregating gene is the same in every case, being a random gene from B_1, or both the maternal and the paternal genes of B_1 are among the v genes. Since

$$\Pr(S_1 = S_2 = ... = S_t) \;=\; 2^{-v+1},$$

we obtain

$$\begin{aligned}
\psi(B_1, ..., B_1, B_2, ..., B_m) \;&=\; 2^{-v+1}\Big(\psi(B_1, B_2, ..., B_m) \;+ \\
&\qquad (2^{v-1} - 1)\;\psi(M_{B_1}, F_{B_1}, B_2, ..., B_m)\Big) \\
&=\; 2^{-v}\Big(\psi(M_{B_1}, B_2, ..., B_m) \;+\; \psi(F_{B_1}, B_2, ..., B_m) \\
(3.7)\qquad &\qquad +\; (2^{v} - 2)\;\psi(M_{B_1}, F_{B_1}, B_2, ..., B_m)\Big)\,.
\end{aligned}$$

Together with symmetry and boundary conditions, these equations determine the multiple kinship coefficients on any pedigree. Note that the number of arguments of ψ is never increased by recursion, although the number of terms may be doubled at each step. Practical implementation can therefore be problematic on a large multi-generation pedigree if the initial number m of genes or individuals considered is more than about 7.

The m-gamete kinship coefficients can be used to determine probabilities of patterns of gene *ibd* among a set of m genes. First, however, a specification of such patterns (gene *ibd* states) is needed. Among a set of genes in given individuals, a gene *ibd* state is a partition of the genes into subsets that are *ibd*. We denote such a pattern by **J**, and refer to it as the *pattern* of gene identity by descent among the individuals. A partition of m ordered genes may be specified by a set of m integers as follows. Let $k_1 = 1$. Suppose genes $1, 2, \ldots, r$ have been assigned v distinct labels $k_1, \ldots, k_r$. If gene $r + 1$ is *ibd* to a previous gene l, $k_{r+1} = k_l$. Otherwise, $k_{r+1} = v + 1$. (For the case $m = 4$, this labeling is shown in Table 3.1.) As m increases, the number of possible states of gene *ibd* increases rapidly. For the 12 genes of 6 individuals, there are more than 4 million gene identity states (partitions of 12 ordered objects). However, for the analysis of phenotypic data on individuals, one need not distinguish the paternal and maternal genes of an individual. The interchange of labels on the two genes within each member of any subset of the individuals groups the *ibd* states into *genotypically distinct* classes of states. For the case of two individuals, this grouping is also shown in Table 3.1. This grouping substantially decreases the number of patterns of gene *ibd* that must be considered. For example, for six individuals there are only just over 198,000 genotypically distinct classes of states (Thompson, 1974). Although this is not a small number, with modern computers and an efficient indexing of state classes it is not impossible to consider all the possible state classes given data on 6 individuals.

Returning to the relationship between multi-gamete kinship and gene *ibd* state probabilities, consider any specified (detailed or grouped) *ibd* state among the genes of a set of individuals. For example, for five individuals $(B_1, B_2, B_3, B_4, B_5)$ the state $(1,2,1,3,4,4,2,4,2,5)$. This state contributes 0.25 to $\psi(B_1, B_2)$, 0.5 to $\psi(B_3, B_4)$ and 0.125 to $\psi(B_1, B_4, B_5)$. Conversely, any multi-gamete kinship coefficient among individuals, say $\psi(B_1, \ldots, B_m)$ can be written as a weighted sum of *ibd* state probabilities:

$$\psi(B_1, \ldots, B_m) \quad = \quad \sum_{\mathbf{J}} \Pr(\text{segregating genes } ibd \mid \mathbf{J}) \Pr(\mathbf{J}).$$

If multi-gamete kinship coefficients are computed for all subsets of the individuals of interest, the linear equations may be inverted to give the *ibd* state probabilities, $\Pr(\mathbf{J})$ among the genes of the individuals. Karigl (1981) was interested primarily in the determination of the probabilities of patterns of *ibd* among the four genes of two individuals, at a single genetic locus. He gives details of the equations for this case.

3.5 Patterns of gene *ibd* in pairs of individuals

Among the four genes of two individuals at a single autosomal locus, there are 15 states of gene identity (Cotterman, 1974). These are shown in Table 3.1, and correspond simply to the number of partitions of the four genes into classes of genes that are *ibd*. However, there are only 9 groups of genotypically distinct classes of states, since with regard to genotypes the maternal and paternal origins of genes are irrelevant, so the identities of the two genes within each individual can be interchanged. For the case of two individuals, the state classes can be characterized by specifying the autozygous individual(s), and the number of genes shared *ibd* between the two individuals (Table 3.1).

For two non-inbred diploid individuals, there are only three possible genotypically distinct *gene identity states* at a single autosomal locus. That is, the individuals can share neither of their genes *ibd*, or one, or both. These events have probabilities $\kappa = (\kappa_0, \kappa_1, \kappa_2)$ say, $(\kappa_0 + \kappa_1 + \kappa_2 = 1)$, determined by the pedigree. Individuals are related if $\kappa_0 < 1$. Each relationship may thus be represented by a point in an equilateral triangle of unit height, the vertices corresponding to unrelated pairs $(\kappa_0 = 1)$, parent-offspring $(\kappa_1 = 1)$, and the identity (monozygous twin) relationship $(\kappa_2 = 1)$. (Care should be taken in applying the standard equations to monozygous twins, since they result from a single maternal and a single paternal meiosis.) The triangle representation is shown in Figure 3.2 and

ibd pattern		*ibd* label	*ibd* group	state description	
B_1	B_2			individuals	genes
p m	p m			autozygous	shared
● ●	● ●	1 1 1 1	1 1 1 1	B_1, B_2	4 genes *ibd*
● ●	● ○	1 1 1 2	1 1 1 2	B_1	3 genes *ibd*
● ●	○ ●	1 1 2 1			
● ○	● ●	1 2 1 1	1 2 1 1	B_2	3 genes *ibd*
● ○	○ ○	1 2 2 2			
● ●	○ ○	1 1 2 2	1 1 2 2	B_1, B_2	none
● ●	○ †	1 1 2 3	1 1 2 3	B_1	none
● ○	† †	1 2 3 3	1 2 3 3	B_2	none
● ○	● ○	1 2 1 2	1 2 1 2	none	2 genes
● ○	○ ●	1 2 2 1			shared
● ○	● †	1 2 1 3	1 2 1 3	none	1 gene
● ○	† ●	1 2 3 1			shared
● ○	○ †	1 2 2 3			
● ○	† ○	1 2 3 2			
● ○	† ⋆	1 2 3 4	1 2 3 4	none	none

Table 3.1. *States of gene ibd among the four genes of two individuals*

the values of κ for some standard relationships are give in Table 3.2. The kinship coefficient is the probability that homologous genes segregating from two individuals are identical by descent and thus $\psi = (2\,\kappa_2 + \kappa_1)/4$. Lines of constant kinship are orthogonal to the line $\kappa_1 = 0$. Sibs, with $\kappa = (1/4, 1/2, 1/4)$ have the same kinship coefficient as a parent-offspring relationship. Half-sibs, with $\kappa = (1/2, 1/2, 0)$ have the same kinship coefficient as double-first-cousins ($\kappa = (9/16, 3/8, 1/16)$).

Pairwise relationship	κ_0	κ_1	κ_2	ψ
Unrelated	1.00	0	0	0
Parent-offspring	0	1.00	0	0.25
Monozygous twin	0	0	1.00	0.50
Full Sib	0.25	0.50	0.25	0.25
Half sib, grandparent, aunt	0.50	0.50	0.00	0.125
First cousin	0.75	0.25	0	0.0625
Double first cousin	0.5625	0.375	0.0625	0.125
Quadruple half first cousin	0.5312	0.4375	0.0312	0.125

Table 3.2. *Values of κ, and kinship coefficient ψ, for some standard relationships between two non-inbred individuals*

While each relationship determines a point κ, the converse is not true. Several relationships give the same probabilities κ; the simplest example is the three pairwise relationships grandparent-grandchild, half-sibs, and aunt-niece, all of which have $\kappa = (1/2, 1/2, 0)$. Moreover, some points in the triangle are not (even in

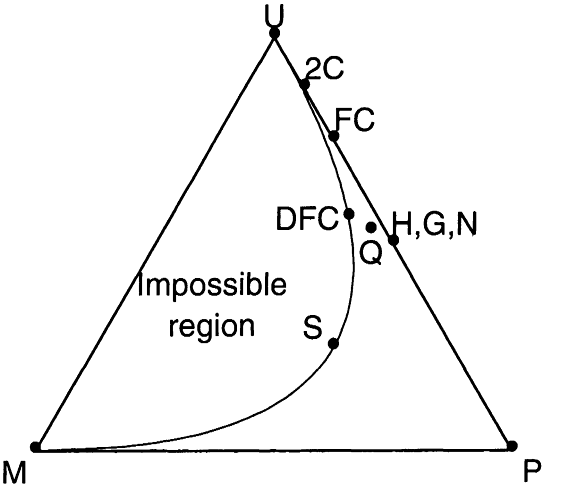

FIGURE 3.2. *The relationship triangle for non-inbred relatives*

the limit) attainable by any relationship. In fact, it can be shown that $\kappa_1^2 \geq 4\kappa_0\kappa_2$ (Thompson, 1986). This result follows from the fact that, for non-inbred individuals

$$\begin{aligned} \psi &= (1/4)(\psi_{MM} + \psi_{FF} + \psi_{MF} + \psi_{FM}) \\ \text{and } \kappa_2 &= (\psi_{MM}\psi_{FF} + \psi_{MF}\psi_{FM}) \end{aligned} \tag{3.8}$$

where the subscripted kinship coefficients are those between a parent (mother (M) or father (F)) of one individual, and a parent of the other. Then the arithmetic-geometric mean inequality gives

$$\begin{aligned} 4\kappa_2 &\leq (\psi_{MM} + \psi_{FF})^2 + (\psi_{MF} + \psi_{FM})^2 \\ &\leq (\psi_{MM} + \psi_{FF} + \psi_{MF} + \psi_{FM})^2 \\ &= (4\psi)^2 = (\kappa_1 + 2\kappa_2)^2 \\ &= \kappa_1^2 + 4\kappa_2(\kappa_1 + \kappa_2) \quad \text{or} \\ 4\kappa_2\kappa_0 &= 4\kappa_2(1 - (\kappa_1 + \kappa_2)) \leq \kappa_1^2. \end{aligned}$$

In order for equality to hold in this inequality, one pair of the crossparental kinship coefficients must be 0, and the other pair equal. Such relationships include full sibs ($\psi_{MM} = \psi_{FF} = 1/4$, $\psi_{MF} = \psi_{FM} = 0$) and double-cousins of any degree v, for which $\psi_{MM} = \psi_{FF} = (1/2)^{v+2}$, $\psi_{MF} = \psi_{FM} = 0$ or $\psi_{MF} = \psi_{FM} = (1/2)^{v+2}$, $\psi_{MM} = \psi_{FF} = 0$. These relationships give values of κ falling on the boundary parabola.

It is possible for the mother and father of each individual to be related to both the mother and the father of the other, without either individual being inbred. That is, all four of the cross-parental kinship coefficients in the above equation may

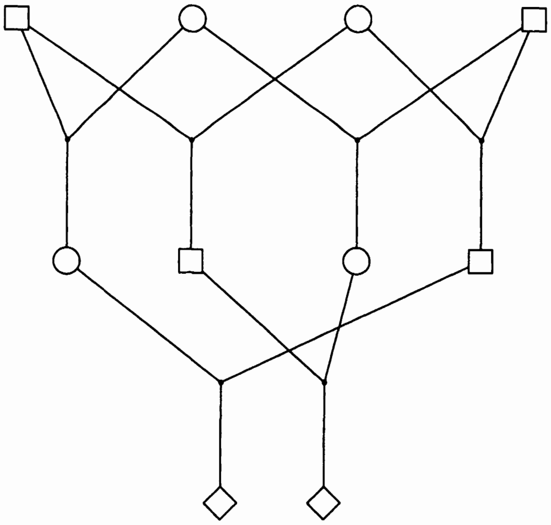

FIGURE 3.3. *The relationship of quadruple-half-first-cousins*

be non-zero. The simplest example is that of quadruple-half-first-cousins, shown in Figure 3.3. For this relationship, the mother and the father of each individual is a half-sib of both the mother and the father of the other, so $\psi_{MM} = \psi_{FF} = \psi_{MF} = \psi_{FM} = (1/8)$. Hence, using equation (3.8), $\kappa_2 = 1/32$, $\kappa_1 = 7/16$, $\kappa_0 = 17/32$ and $\psi = 1/8$. The point in the triangle lies midway between that for half-sibs and for double-first-cousins, which also each have $\psi = 1/8$.

More details of the material of this section, and references to earlier work, can be found in Chapter 2 of Thompson (1986).

3.6 Observations on related individuals

Phenotypic similarities among relatives result from the genes they share *ibd*. Among an ordered set of genes, a partition of the set may be used to specify which subsets of the genes are *ibd* (section 3.4). Again we denote such a pattern of gene *ibd* by **J**. In section 1.3, the meiosis indicators were defined (equations (1.2) and (1.3)), and it was seen how the meiosis indicators $S_{\bullet,j}$ determine descent of founder genes, and patterns of gene identity by descent, at any given locus j. Thus, the passage of genes in pedigrees provides the connection between observable genetic characteristics and the pedigree structure, whether we are estimating relationships from genetic data, estimating the genetic basis of traits knowing the pedigree, or inferring the ancestry and descent of particular genes, knowing both the genetic model and the data (section 1.4).

In particular, we consider a currently observed set of individuals, and the pattern,

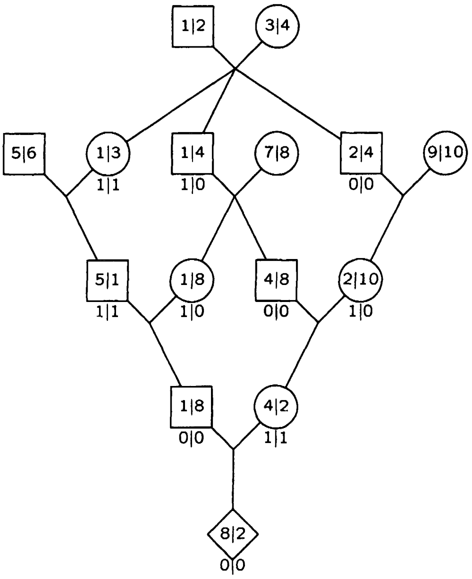

FIGURE 3.4. *Meiosis indicators $S_{\bullet,j}$ determine descent of founder genes, and patterns of gene identity by descent, at any given locus j: see Figure 1.2*

J, of genes *ibd* among them, at a single locus. We therefore drop the locus index j, and write $\mathbf{S} = \{S_i; i = 1, ..., m\}$ (equation (1.1)), for the m meioses of the pedigree. The example of Figure 1.2 is shown again in Figure 3.4. The meiosis indicators shown under each individual are for the paternal and for the maternal meiosis to that individual, respectively. Then **S** determines the pattern, **J**, of genes *ibd* in this currently observed set of individuals; $\mathbf{J} = \mathbf{J}(\mathbf{S})$. The probability of any phenotypic data **Y** (i.e. observed characteristics of the individuals) depends on **S** only through $\mathbf{J}(\mathbf{S})$, and so

$$\text{(3.9)} \qquad \begin{aligned} \Pr(\mathbf{Y}) &= \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{S}) \Pr(\mathbf{S}) \\ &= \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S})) \Pr(\mathbf{S}) \\ &= \sum_{\mathbf{J}} \Pr(\mathbf{Y} \mid \mathbf{J}) \Pr(\mathbf{J}). \end{aligned}$$

Equation (3.9) may be compared with equation (1.5) of Chapter 1. In equation (1.5) the latent variables were the genotypes G_i of individuals, whereas here they are the meiosis indicators. In both cases, the form of the likelihood is that of a latent variable problem (section 2.4), and either may be the more convenient for likelihood computation and inference (Chapter 6).

In partitioning the likelihood as in equation (3.9), the "genetic model" is separated from the effects of genealogical and genetic structure. The probability of a set of meiosis indicators **S** at a single locus is trivial; the components are independent, each 0 or 1 with probability 1/2. The probability of a given pattern **J**(**S**) depends on the genealogical relationship among the observed individuals: in principle it may be computed by the methods of sections 3.4 or 3.8. Given the gene identity pattern, **J**(**S**), the probability of data depends on the different types of genes, their frequencies, and how they affect observable phenotypes.

Now consider the probability $\Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S}))$, for a specified pattern of gene *ibd* among the observed individuals. The probability any distinct gene, k, is of allelic type $a(k)$ is the population frequency, $q_{a(k)}$, of the allele. Distinct genes k have independent allelic types. Thus, $\Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S}))$ is the sum over all possible assignments $\mathcal{A}$ of allelic types to genes of the product of allele frequencies $q_{a(k)}$ of assigned alleles $a(k)$:

$$\Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S})) \;=\; \sum_{\mathcal{A}} \prod_{k} q_{a(k)}. \tag{3.10}$$

This equation was given by Thompson (1974) who gave an example of *ABO* blood types on three individuals. The special case of two individuals (9 states **J**) is discussed in Chapter 2 of Thompson (1986).

In general, efficient determination of all allocations $\mathcal{A}(j)$ at locus j compatible with data Y_{*j} is straightforward for genotypic data (for example, DNA marker phenotypes). An algorithm for this determination of is given by Kruglyak et al. (1996). The implementation we use is due to Simon Heath (personal communication) and is described in more detail by Thompson and Heath (1999). We use the same example pedigree, with the values of $S_{\bullet,j}$ given in Figure 1.2, and assume five individuals observed with the genotypes shown in Figure 3.5(a). The method rests first on the fact that only founder genes having copies in observed individuals are constrained in allelic type: in our example, the genes labeled $\{1, 2, 4, 5, 8, 10\}$. Further two genes constrain each other's allelic type only when both are present in an observed individual. The *gene graph* (Figure 3.5(b)) connects all such pairs of genes. Allocation of allelic types may be considered separately for each component subgraph of connected genes. In our example, the genes $\{1, 5\}$ may be considered independently of $\{2, 4, 8, 10\}$. This assignment is readily accomplished, even on a much larger example. For given $S_{\bullet,j}$ there are in general only 2, 1 or 0 possible assignments of allelic types to the genes of a component subgraph. For our example, there are two possible assignments for the first component and one for the second: $(a(1), a(5)) = (A, C)$ or (C, A) and $(a(2), a(4), a(8), a(10)) = (C, D, C, B)$. The algorithm can in principle be generalized to more complex phenotypes, using the conditional independence structure of the gene graph (Figure 3.5(b)), but the procedure becomes far more computationally intensive.

For completeness, and as an example of the above general formula, consider again the case of a non-inbred pair of relatives. There are then three *ibd* states J_0, J_1 and

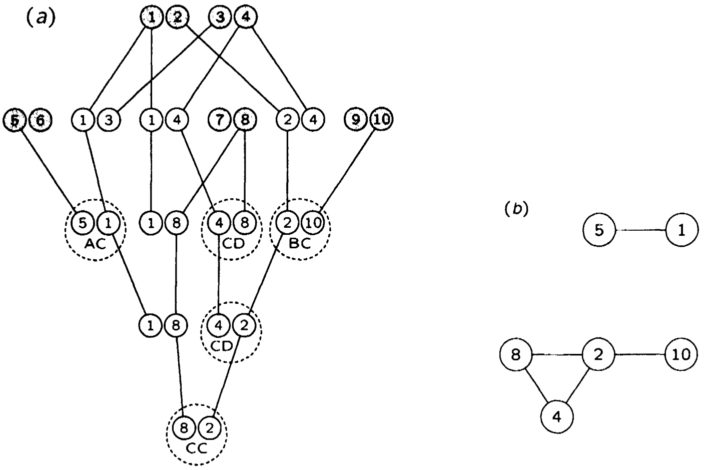

FIGURE 3.5. *Determination of probabilities* $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$. *The gene descent pattern is assumed to be that of Figure 1.2, and the pairs of genes are shown, rather than the individuals. Five individuals, shown as dashed circles, are assumed to be observed, with marker genotypes as indicated: see text for details. (a) Only genes present in observed individuals are constrained in type. (b) Two genes in a single observed individual are jointly constrained*

J_2, with probabilities κ_0, κ_1, and κ_2, these being determined by the relationship R between the individuals (section 3.5). The state J_k denotes that k genes are shared *ibd* between the two individuals. Suppose at a given locus the ordered genotypes of the pair are (G_1, G_2). Then the analogue of equation (3.9) is

$$\begin{aligned}\Pr(G_1,G_2;R) &= \kappa_0(R)\Pr(G_1,G_2;J_0) + \kappa_1(R)\Pr(G_1,G_2;J_1) \\ &\qquad + \kappa_2(R)\Pr(G_1,G_2;J_2).\end{aligned}$$

Now, $\Pr(G_1,G_2;J_2) = \Pr(G_1)$, the population frequency of the genotype, if $G_1 = G_2$, and 0 otherwise. This is the probability for a pair of monozygous twins. Also, $\Pr(G_1,G_2;J_0) = \Pr(G_1)\Pr(G_2)$, the probability for an unrelated pair of relatives. Finally, $\Pr(G_1,G_2;J_1)$ is the probability for a parent-offspring pair; these probabilities were given in Table 2.1 (section 2.3). For a pair of relatives, in most cases equation (3.10) take form too trivial to be illuminating. The one non-trivial case is $\Pr(G_1 = A_1A_2, G_2 = A_1A_2; J_1)$. Here the *ibd* gene may be either the A_1 or the A_2 allele; there are two feasible allocations $\mathcal{A}$ of allelic types to the three distinct genes in the two individuals (A_1, A_2, A_1) or (A_1, A_2, A_2) giving a total probability $p_1p_2p_1 + p_1p_2p_2 = p_1p_2(p_1 + p_2)$ as given in the Table 2.1.

Thus, to obtain the probability of genotypes (and hence of phenotypes) for any pair of non-inbred relatives, it is enough to know the probabilities for monozygous-twin, parent-offspring, and unrelated pairs. For a general pair of relatives, however,

Gene *ibd* state	Prior (pedigree) probability	
for two sibs with	(a) an aunt	(b) a niece or half-sib
Sibs sharing 2 *ibd*		
1 2 1 2 1 3	1/8	1/8
1 2 1 2 3 4	1/8	1/8
Sibs sharing 1 *ibd*		
1 2 1 3 1 4	1/8	1/8
1 2 1 3 2 3	1/16	0
1 2 1 3 2 4	1/16	1/8
1 2 1 3 3 4	1/16	1/8
1 2 1 3 4 5	3/16	1/8
Sibs sharing 0 *ibd*		
1 2 3 4 1 3	1/16	0
1 2 3 4 1 5	1/16	1/8
1 2 3 4 3 5	1/16	1/8
1 2 3 4 5 6	1/16	0

TABLE 3.3. *Gene ibd state probabilities at a single locus for a pair of sisters with an aunt, niece, or half-sib. The states are given in the reduced genotypic state-class form, in which the paternal and maternal genes of the three individuals are not distinguished*

the nine genotypically distinct *ibd* patterns of Table 3.1 are required. The probabilities of the states must be computed (see, for example, Karigl (1981)), and also the probabilities of genotypes under each *ibd* state. Again, the latter are special cases of equation (3.10), and are given by Thompson (1986).

Finally, in this section, note that joint analysis of data on a set of relatives is always more powerful than pairwise analysis. A simple example which derives from an actual study is that of a pair of full sibs and their aunt or niece (Browning and Thompson, 1999). Due to the symmetry of the pairwise aunt-niece relationship, pairwise analysis cannot distinguish these relationships; nor distinguish the possibility that the third individual is a half-sister to the pair of sibs. However, two sibs and a aunt can carry six distinct genes at a locus, but sibs with a niece or half-sib cannot. The probabilities of the *ibd* states among the three individuals at a single locus are shown in Table 3.3. Loci at which the two sibs share both their genes *ibd* give the same probabilities of sharing with the third individual under the three possibilities of aunt, niece or half-sib. However, the other state probabilities differ, with the greatest power to distinguish an aunt from a niece or half-sib coming from those loci at which the full sibs do not share any genes *ibd*. Note that data at unlinked loci remains insufficient to distinguish the possibilities that the third individual is a niece or half-sister. As will be seen in section 4.5, data at linked loci results in identifiability of these two alternatives.

3.7 Monte Carlo estimation of expectations

Although the methods of section 3.4 are easily implemented, where large numbers of individuals are considered jointly computation may become impractical or even infeasible. Where exact probabilities cannot be computed, Monte Carlo estimation is an alternative. We use this section to introduce some important ideas in the Monte Carlo estimation of sums, integrals, or expectations. We shall use these methods to estimate probabilities of gene *ibd* patterns in section 3.8. These methods will be important for Chapters 7 and 8. Since, in this section, the latent variables are general, we use the notation **X** instead of **S**.

To estimate $\sum_{\mathbf{x}} g(\mathbf{x})$ the sum may be written as an expectation

$$\sum_{\mathbf{x}} g(\mathbf{x}) \quad = \quad \sum_{\mathbf{x}} \frac{g(\mathbf{x})}{\Pr(\mathbf{X}=\mathbf{x})} \Pr(\mathbf{X}=\mathbf{x}) \quad = \quad \mathrm{E}_P \left(\frac{g(\mathbf{X})}{P(\mathbf{X})} \right)$$

where $P(\cdot)$ is some distribution over $\mathcal{X}$, the space of values of **X**. The distribution $P(\cdot)$ must assign positive probability to every value **x** of **X** for which $g(\mathbf{x}) > 0$. If $X^{(1)}, ..., X^{(N)}$ are simulated from the distribution $P(\cdot)$,

$$\frac{1}{N} \sum_{\tau=1}^{N} \left(\frac{g(\mathbf{X}^{(\tau)})}{P(\mathbf{X}^{(\tau)})} \right) \tag{3.11}$$

is an unbiased estimator of the sum $\sum_{\mathbf{x}} g(\mathbf{x})$. Of course, it may not be a very good estimator; in fact, it may be a very bad estimator. The art of Monte Carlo is finding good distributions to simulate from, and good ways of simulating from them, in order to get good estimators. A "good" estimator is one with small variance. Note this is not the standard statistical paradigm where parameters are estimated from data. In that case, variances are over (hypothetical) repetitions of the experiment or random process giving rise to the data. In Monte Carlo, the relevant variances are Monte Carlo variances.

The simplest form of Monte Carlo is where we simulate independent, identically distributed realizations from some distribution $P(\cdot)$. Note that any sum of terms $g(\mathbf{x})$ is an expectation of $g^*(\mathbf{X}) = g(\mathbf{X})/P(\mathbf{X})$ with respect to the probability distribution $P(\cdot)$. The estimator (3.11) is then an average of terms $g^*(\mathbf{X})$, and, for independent realizations, the Monte Carlo variance of this estimator is

$$N^{-1} \left(\mathrm{E}_P((g^*(\mathbf{X}))^2) \quad - \quad (\mathrm{E}_P(g^*(\mathbf{X})))^2 \right)$$

$$\text{or} \quad \left(\sum_{\mathbf{x}} (g^*(\mathbf{x})^2 P(\mathbf{x})) \quad - \quad \left(\sum_{\mathbf{x}} g^*(\mathbf{x}) P(\mathbf{x}) \right)^2 \right)$$

which, substututiting $g^*(\mathbf{x}) = g(\mathbf{x})/P(\mathbf{x})$, is

$$N^{-1} \left(\sum_{\mathbf{x}} \left(\frac{g(\mathbf{x})^2}{P(\mathbf{x})} \right) \quad - \quad \left(\sum_{\mathbf{x}} g(\mathbf{x}) \right)^2 \right).$$

This may be estimated by the sample variance from the Monte Carlo:

$$(N(N-1))^{-1}\left(\sum_{\tau=1}^{N}\left(g^*(\mathbf{X}^{(\tau)})\right)^2 - N^{-1}\left(\sum_{\tau=1}^{N}\left(g^*(\mathbf{X}^{(\tau)})\right)\right)^2\right)$$

$$\text{or}\quad (N(N-1))^{-1}\left(\sum_{\tau=1}^{N}\left(\frac{g(\mathbf{X}^{(\tau)})}{P(\mathbf{X}^{(\tau)})}\right)^2 - N^{-1}\left(\sum_{\tau=1}^{N}\left(\frac{g(\mathbf{X}^{(\tau)})}{P(\mathbf{X}^{(\tau)})}\right)\right)^2\right).$$

On pedigrees, the simplest distribution to simulate from is the prior distribution on genotypes, which is done by "gene dropping". Genes are assigned to the founders of the pedigree, segregation of genes down the pedigree is simulated, and the required statistics relating to the resultant current genes are computed. Such Monte Carlo estimates have been used by Edwards (1967) to estimate inbreeding coefficients, by MacCluer et al. (1986) to study the loss of genes in pedigrees of endangered species, and by Thompson et al. (1978) to study the potential power of a pedigree study.

Using equation (3.11) is often ineffective. Methods of more effective simulation normally involve some form of *importance sampling.* Note that

$$\begin{aligned} \mathrm{E}_P(g^*(\mathbf{X})) &= \sum_{\mathbf{x}} g^*(\mathbf{x})\,P(\mathbf{x}) \\ &= \sum_{\mathbf{x}} g^*(\mathbf{x})\frac{P(\mathbf{x})}{P^*(\mathbf{x})}\,P^*(\mathbf{x}) \\ &= \mathrm{E}_{P^*}\left(g^*(\mathbf{X})\frac{P(\mathbf{X})}{P^*(\mathbf{X})}\right) \end{aligned} \tag{3.12}$$

provided

$$P^*(\mathbf{X}) > 0 \quad \text{if} \quad g^*(\mathbf{X})P(\mathbf{X}) > 0. \tag{3.13}$$

Thus realizations from $P^*(\cdot)$ can be reweighted in order to estimate expectations under P. Where this is done in such a way that terms making larger contributions to the sum are realized with larger probabilities, this is *importance sampling.* Such sampling decreases the Monte Carlo variance of the estimator of the sum. The effectiveness of this approach depends on the choice of $P^*(\cdot)$. It works best when the summand $g^*(\mathbf{X})\ P(\mathbf{X})$ is the "same shape" as $P^*(\mathbf{X})$, since then the variance of $g^*(\mathbf{X})\ P(\mathbf{X})/P^*(\mathbf{X})$ is small. Ideally, if $P^*(\mathbf{X}) \propto g^*(\mathbf{X})P(\mathbf{X})$, the variance of $g^*(\mathbf{X})\ P(\mathbf{X})/P^*(\mathbf{X})$ is zero. However, this would mean

$$P^*(\mathbf{X}) = \frac{g^*(\mathbf{X})P(\mathbf{X})}{\sum_{\mathbf{x}} g^*(\mathbf{x})\ P(\mathbf{x})} = \frac{g(\mathbf{X})}{\sum_{\mathbf{x}} g(\mathbf{x})},$$

and if the denominator were known the Monte Carlo would be pointless! (Hammersley and Handscomb, 1964). The "same shape" criterion is most important in the tails of the distribution $P^*(\cdot)$; it is a problem if $P^*(\mathbf{X})$ is very small when $g^*(\mathbf{X})P(\mathbf{X})$ is not, since then with low probability there will be very large terms in the estimator, and the Monte Carlo variance will be high. In order to be able to use a given $P^*(\cdot)$ we need first to be able to simulate from it, and second to compute $g^*(\mathbf{x})P(\mathbf{x})/P^*(\mathbf{x})$ at the realized values $\mathbf{x}$ of $\mathbf{X}$. This is sometimes far from straightforward, but we defer further discussion to Chapter 7.

Note the difference between a "simulation study" and a "Monte Carlo analysis". Simulation studies are typically undertaken to discover empirically the distribution of a test statistic, or to assess the potential power of a study design. It involves the simulation of data random variables under a model of interest. In a Monte Carlo analysis, integrals, sums, or expectations are estimated by simulating random variables from some distribution, but the random variables are not normally the data random variables (often, the data are fixed) and the distribution is simply a tool to provide good estimates of the required expectations. In practice, the difference may be slight. The probability distribution we simulate from in a Monte Carlo estimation problem may often be closely related to the probability model underlying the data in a statistical problem. Conversely, the probability distribution we use in a simulation study could equally be a convenient tool, with reweighting used to adjust the realizations to the distribution of interest (equation (3.12)). In a Monte Carlo analysis we shall normally simulate conditional on fixed data, but in a simulation study it may sometimes also be desirable to simulate potential data conditional on partial data already obtained.

3.8 Reduction of Monte Carlo variance

The earliest use of Monte Carlo estimation on pedigrees was to estimate inbreeding coefficients. Before digital computers were available, Wright and McPhee (1925) traced random paths up pedigrees. By random choice of a male or female parent, one is realizing the ancestry of a particular allele, and hence realizations of the *ibd* status of, for example, the two genes within a current individual. Much more recently, using a computer, Edwards (1967) realized the descent of genes down pedigrees to estimate inbreeding coefficients. In effect, both Wright and McPhee (1925) and Edwards (1967) are realizing latent variables $\mathbf{S}$. To estimate the probability of a specified *ibd* pattern, $\mathbf{J}^*$, define

$$(3.14) \qquad \begin{aligned} g^*(\mathbf{S}) &= 1 \text{ if } \mathbf{J}(\mathbf{S}) = \mathbf{J}^* \\ &= 0 \text{ otherwise.} \end{aligned}$$

Then the probability of the pattern $\mathbf{J}^*$ is the expectation of $g^*(\mathbf{S})$ under the distribution of the random descent of genes in pedigrees.

Any probability can be estimated as the expectation of an indicator variable in this way, but the method is often not very efficient, if only the probability of a particular $\mathbf{J}^*$ is needed. On the other hand, if the probabilities of all *ibd* patterns among a given set of current genes are desired, this may be an effective approach; each realization of $\mathbf{S}$ contributes to some *ibd* pattern $\mathbf{J}(\mathbf{S})$. Different realizations $\mathbf{S}$ are, of course, independent, but the probabilities of different *ibd* patterns $\mathbf{J}$ estimated from the same set of realizations are dependent. It is important to recognize this dependence, but it is seldom a practical problem; multinomial covariances are small for large Monte Carlo samples.

Another key idea in effective Monte Carlo is "Rao-Blackwellization" of estimators. This procedure is named for the classic Rao-Blackwell Theorem in Statistics, whereby the statistical variance of an estimator $g(\mathbf{X})$ is reduced by replacing it by its conditional expectation given some statistic T: if $h(T) = \mathrm{E}(g(\mathbf{X})|T)$,

$$\mathrm{E}(h(T)) \;=\; \mathrm{E}(g(X)) \;\text{ and }\; \mathrm{var}(h(T)) \leq \mathrm{var}(g(\mathbf{X})).$$

Here we replace a part of the Monte Carlo by exact computation of a (conditional) probability or expectation. Formally, suppose the latent variables $\mathbf{X}$ are divided into two sets of components $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. As before, we wish to estimate $\mathrm{E}_P(g^*(\mathbf{X})) = \mathrm{E}_P(g^*(\mathbf{X}_1, \mathbf{X}_2))$, where each of $\mathbf{X}_1$ and $\mathbf{X}_2$ is a (possibily vector) variable. If pairs $(\mathbf{X}_1^{(\tau)}, \mathbf{X}_2^{(\tau)})$, $i = 1, ..., N$ are independently realized from the probability distribution $P(\cdot)$, one estimator is (see equation (3.11))

$$T_N^* \;=\; \frac{1}{N}\sum_{i=1}^{N} g^*(\mathbf{X}_1^{(\tau)}, \mathbf{X}_2^{(\tau)}).$$

Suppose it is possible to compute $h(\mathbf{X}_1) = \mathrm{E}_P(g^*(\mathbf{X}_1, \mathbf{X}_2) \mid \mathbf{X}_1)$. Another Monte Carlo estimator is then

$$T_N = \frac{1}{N}\sum_{i=1}^{N} h(\mathbf{X}_1^{(\tau)}).$$

Then the Monte Carlo variance of T_N is easily shown to be no larger than that of T_N^*, and usually strictly smaller. Whether such Rao-Blackwellization is computationally effective depends on whether the increased cost of computing $h(\mathbf{X}_1)$ rather than $g^*(\mathbf{X}_1, \mathbf{X}_2)$ is outweighed by the reduction in the number of the Monte Carlo realizations required to achieve a given precision. There is no general rule; see also section 9.4.

Returning to realizations of gene descent in pedigrees, suppose we wished to estimate by Monte Carlo the inbreeding coefficient of the offspring of double first cousins: in fact, the answer is 0.125 (Table 3.2). If we use the estimator of equation (3.14), scoring 1 for each realization of $\mathbf{S}$ in which the final offspring individual is autozygous (has two *ibd* genes), the Monte Carlo variance is that of a binomial proportion for probability 1/8: $(1/8)(7/8)(1/N) \;=\; 0.1094/N$. If instead, we score *ibd* patterns in the double-first cousins, we have a trinomial realization of $\boldsymbol{\kappa} \;=\; (\kappa_0, \kappa_1, \kappa_0) \;=\; (9/16, 6/16, 1/16)$. Then the inbreeding coefficient of the offspring is estimated by $\widehat{\psi} = (2\widehat{\kappa_2} + \widehat{\kappa_1})/4$, which has Monte Carlo variance

$$\begin{aligned}(1/4)\mathrm{var}(\widehat{\kappa_2}) + (1/4)\mathrm{cov}(\widehat{\kappa_2}, \widehat{\kappa_1}) + (1/16)\mathrm{var}(\widehat{\kappa_1}) \;&=\; \\ N^{-1}(0.01465 - 0.00586 + 0.01465) \;&=\; 0.02344/N\end{aligned}$$

which is almost 5 times smaller. In this case, $\mathbf{S}_1$ corresponds to the meioses down to the double first cousins, and $\mathbf{S}_2$ to the meioses from the double-first-cousins to their offspring. The original estimator scores 1 or 0 depending on whether or not $(\mathbf{S}_1, \mathbf{S}_2)$

implies autozygosity of the offspring individual. The conditional expectation $h(\mathbf{S}_1)$ is simply the probability of autozygosity in the final individual, given the particular $\mathbf{S}_1$ realized.

As another example, consider estimation of the inbreeding coefficient of the final individual of the pedigree of Figure 3.1. The actual value is $7/64 = 0.1094$ (section 3.2), so using direct gene-drop, the Monte-Carlo standard error is $\sqrt{(7/64)(57/64)/N} = 0.3121/\sqrt{N}$. Alternatively, we may use Monte Carlo only to the parents of the individual. In this case, there are nine possible states of gene *ibd* among four genes of these two parent individuals (Table 3.1). For each *ibd* pattern in the parents, the conditional expectation of the indicator of autozygosity of the offspring is simply the conditional probability given the parental *ibd* state. These probabilities that the final individual, B, receives two *ibd* genes, range from 1.0, for the parental pattern 1111, down to 0.0 for the pattern 1234:

$$\begin{aligned} f_B = \psi(M_B, F_B) &= \Pr(1111) + 0.5(\Pr(1112) + \Pr(1121) + \Pr(1211) \\ &\quad + \Pr(1222) + \Pr(1212) + \Pr(1221)) + \\ &\quad 0.25(\Pr(1213) + \Pr(1231) + \Pr(1223) + \Pr(1232)). \end{aligned}$$

Here $\Pr(k_1k_2k_3k_4)$ is the probability of that pattern among the four parental genes, the first two being the genes of one parent and the last two of the other. The Monte Carlo standard error of this estimate is approximately $0.17/\sqrt{N}$, in this case estimated empirically. There are two sources in the gain in efficiency, one replacing a part of the Monte-Carlo by an exact computation of an expectation (Rao-Blackwellization), and second the negative covariances of the Monte-Carlo multinomial proportions providing the estimated *ibd* pattern probabilities in the parents. Since $\psi(M, F)$ is a positive linear combination of these *ibd* pattern probabilities, the negative covariance reduces the Monte Carlo standard error of the estimate of $\psi(M, F)$. This idea is a little different from the use of *antithetic* variates (Hammersley and Handscomb, 1964), but of similar effect. Antithetic variates are negatively correlated realizations used to reduce the variance of a sum or average. Here the realizations of $\mathbf{S}$ are independent, but the component events are negatively correlated.