Part III

Human Genetics

9

Mapping Qualitative Traits in Humans Using Affected Sib Pairs

In humans we cannot create inbred lines, backcrosses, etc. Consequently, it is more difficult to study directly the correlation of phenotypes and genetic markers. We can proceed indirectly by noting that relatives frequently have more similar phenotypes than non-relatives, presumably because they have more similar genotypes. For studying human diseases, particularly convenient units are *affected sib pairs* (ASP), which are the subject of this chapter. We delay until Chap. 11 a discussion of the substantially more complex problem of pedigrees involving variable numbers and relationships of affecteds.

Since humans are members of populations, not subject to breeding experiments, we shall want to use some of the material on population genetics from Chap. 3, notably that concerned with the ideas of random mating/Hardy-Weinberg equilibrium and identity by descent (IBD).

If two relatives are affected with the same disease, which is caused to some extent by the individual's genotype and is relatively rare in the population, it seems plausible to hypothesize that they have the disease because both inherited one or more disease-predisposing alleles from a common ancestor.

Recall that two relatives are said to have inherited an allele identical by descent (IBD) at a given locus, if they have inherited the same allele from a common ancestor. At any genetic locus, two siblings inherit their paternal allele IBD with probability 1/2 and independently inherit their maternal allele IBD with probability 1/2. Thus they inherit 0, 1, or 2 alleles IBD with probabilities 1/4, 1/2, 1/4, respectively; and on average they inherit one allele IBD. (This argument presupposes that the parents are not inbred, so they do not already contain alleles inherited IBD from a remote ancestor.) If we have a sample of, say, n sib pairs, at a randomly selected genetic marker, there will be about n alleles inherited IBD. If the sib pairs share a given phenotype, e.g., the same disease, then we expect that at a marker tightly linked to a gene or genes contributing to the phenotype there will be more than n alleles IBD. In the following sections we develop a genetic model for a qualitative trait and discuss genome scans to detect genes contributing to the trait. For

simplicity we refer to the trait as a disease and individuals having the disease as affected. Analysis of QTL in humans is discussed in Chap. 11.

We use a simple genetic model in order to describe a potential connection between the disease status of the siblings and the distribution of the number of alleles shared IBD. This connection is then used in order to derive the conditional distribution of the number of alleles shared IBD, given that both siblings are affected. This leads in turn to a relation that connects the frequency in the population of the susceptibility alleles and their contribution to the risk of getting the disease to the distribution of the test statistics. These issues are discussed in the following two sections.

The third section deals with the asymptotic distribution of the test statistic, and the properties of the associated test, when large samples are used in order to detect a risk factor that has a relatively small effect on the probability of being affected.

The IBD status for a pair at a given locus is inferred from the genotypic information at hand, which may include the genotypes of the siblings and their parents or the genotypes of the siblings alone. In the preceding discussion we make the assumption that the IBD status can be perfectly reconstructed based on the genotypic information. In practice, this is seldom the case and an estimate of the IBD number has to replace the unknown true value. In the fourth section we will present statistical tools for the estimation of the IBD state from the genotypic information and assess the effect of partial information on the statistical properties of the scanning procedure.

9.1 Genetic Models

We assume initially a single susceptibility locus. The polymorphism at that locus consists of two alleles – a susceptibility allele D, and wild-type allele d. The genetic model provides the ingredients which are needed in order to compute the conditional distribution of the IBD status, given the phenotypes of the siblings. It consists of two components: a model connecting phenotypes to genotypes at the susceptibility locus and a population genetic model, describing the population joint distribution of genotypes at the trait loci for the parents of the siblings. Although we discuss in detail the case of a bi-allelic disease locus, essentially all the results described are valid for loci having an arbitrary number of alleles.

A Model for the Trait

Here we consider a single autosomal trait locus with allele D, associated with the disease, and a wild-type allele d. In the model we allow for sporadic cases and partial penetrance. Specifically, define the three *penetrance* probabilities:

$$g_0 = \Pr(\text{Affected} \,|\, dd) \,,$$
$$g_1 = \Pr(\text{Affected} \,|\, Dd) \,,$$
$$g_2 = \Pr(\text{Affected} \,|\, DD) \,.$$

In order to emphasize certain similarities with the models used for experimental genetics, it is often convenient to re-write the penetrances in the form

$$g_1 = g_0 + \alpha + \delta; \quad g_2 = g_0 + 2\alpha \,.$$

We assume that $\alpha > 0$ to be consistent with $g_2 > g_0$. With this notation y, which refers here to the *probability* that an individual is affected as a function of genotype, has the form of equation (2.2) with $e = 0$. Now x_{M} (resp. x_{F}) equals the number of D alleles, 0 or 1, inherited from the mother (resp. father). Note, however, that while in Chap. 2 y is an observed quantitative phenotype that is allowed to take on any value, here the phenotype is 0 (no disease) or 1 (disease), while y is the *unobserved* (conditional on the genotype) probability that the phenotype is 1. Hence y itself must be between 0 and 1.

The *additive model*, which we emphasize in what follows, relates the three penetrance parameters to each other by requiring that $g_1 = (g_0 + g_2)/2$, or equivalently that $\delta = 0$. Thus, one needs to specify only g_0 and α. Moreover, as we shall see later, the quantity of interest will depend only on the ratio $R_2 = g_2/g_0 = 1 + 2\alpha/g_0$ – the relative risk of a *risk-allele* homozygote with respect to a *wild-type* homozygote. One can envision other special cases of this model, even in the simple case of a single trait locus. Important examples are the *recessive model* which assumes $g_0 = g_1 < g_2$, equivalently $\delta = -\alpha$ or the *dominant model*, which assumes $g_0 < g_1 = g_2$, or $\delta = \alpha$. The statistic we consider below, which counts the total number of alleles shared IBD, is most appropriate for the additive model and to a good approximation for a dominant model as well.

We complete the description of the relation between the genotypes and the probability that the siblings are affected by adding the assumption that within a pedigree, the phenotypes of relatives are conditionally independent given their genotypes. As a result, we find that

$$\Pr(\text{Both affected} \,|\, G_1, G_2) = \Pr(\text{Affected} \,|\, G_1) \times \Pr(\text{Affected} \,|\, G_2) = y_1 y_2 \,, \tag{9.1}$$

where G_1 and G_2 are the genotypes of the first and the second sibling respectively. An important consequence of this assumption is the exclusion of environmental effects on susceptibility to the disease. (See Prob. 9.5 for possible generalizations.)

A Population Genetic Model

The second component in the genetic model is a population genetic model that describes the frequencies of the pedigree founders' genotypes. For the case of

a sib-pair, there are two founders – the mother and the father, whom we assume mate at random in an infinitely large population and are themselves the product of random matings. Hence their genotypes are in Hardy-Weinberg equilibrium, i.e., the two alleles at a given locus are randomly sampled from the population pool. If the population frequency of the allele D is denoted by p (and the frequency of the allele d is $1-p$), then the probability of the genotype DD is p^2. Likewise, the probability of the genotype dd is $(1-p)^2$, and the probability of the genotype Dd is $2\,p(1-p)$. Random mating also implies independence between the parents' genotypes. For example, the probability that both parents' genotypes are DD is p^4. In a similar fashion, one can compute the probability of all other combinations of parents' genotypes as a function of a single parameter p, the frequency of the allele D in the genetic pool. It also follows that each child individually has a genotype that satisfies Hardy-Weinberg frequencies. However, the genotypes of two children are dependent.

From (2.2), we obtain (2.3), which for convenience we repeat here (with $e=0$):

$$y = m + \{\alpha + (1-2p)\delta\}[(x_{\mathrm{M}} - p) + (x_{\mathrm{F}} - p)] - \{2\delta\}[(x_{\mathrm{M}} - p)(x_{\mathrm{F}} - p)] \; .$$

Combining this expression with the assumption of Hardy-Weinberg equilibrium, we also obtain the variance decomposition (2.5):

$$\sigma_y^2 = \sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{D}}^2 \; ,$$

where $\sigma_{\mathrm{A}}^2 = 2p(1-p)[\alpha + (1-2p)\delta]^2$ and $\sigma_{\mathrm{D}}^2 = 4p^2(1-p)^2\delta^2$. For a dominant trait ($\delta = \alpha$) $\sigma_{\mathrm{A}}^2 = 8p(1-p)^3\alpha^2$, while for a recessive trait ($\delta = -\alpha$) $\sigma_{\mathrm{A}}^2 = 8p^3(1-p)\alpha^2$. In the usual case that p is substantially less than $1/2$, the additive variance is much larger than the dominance variance for a dominant trait, smaller for a recessive trait. The simplest case is an additive trait, for which $\delta = 0$, hence $\sigma_{\mathrm{D}}^2 = 0$.

By taking expectations in (9.1) and using the representation of y_i given above to obtain an expression for the product $y_1 y_2$, we can calculate the probability $\Pr(A)$ that two sibs are both affected:

$$\Pr(A) = \mathrm{E}(y_1 y_2) = m^2 + \mathrm{cov}(y_1, y_1) = m^2 + \sigma_{\mathrm{A}}^2/2 + \sigma_{\mathrm{D}}^2/4 \; . \tag{9.2}$$

To see how (9.2) is derived, let $x_{\mathrm{M}i}$ ($x_{\mathrm{F}i}$) denote the number of D alleles inherited by the ith sib from their mother (father). First recall that $\mathrm{E}[(x_{\mathrm{M}i} - p)^2] = p(1-p)$. Now consider the product $(x_{\mathrm{M}1} - p)(x_{\mathrm{M}2} - p)$. If $x_{\mathrm{M}1}$ and $x_{\mathrm{M}2}$ are IBD, then the product equals $(x_{\mathrm{M}1} - p)^2$, so in this case the expected product is just $p(1-p)$, as before. If $x_{\mathrm{M}1}$ and $x_{\mathrm{M}2}$ are not IBD, then by the Hardy-Weinberg assumption, they are independent and the expected product is the product of the expectations, which equals 0. Since $x_{\mathrm{M}1}$ and $x_{\mathrm{M}2}$ are IBD with probability $1/2$, we find that $\mathrm{E}(x_{\mathrm{M}1} - p)(x_{\mathrm{M}2} - p) = p(1-p)/2 + 0/2 = p(1-p)/2$. Similarly $\mathrm{E}[(x_{\mathrm{M}1} - p)(x_{\mathrm{M}2} - p)(x_{\mathrm{F}1} - p)(x_{\mathrm{F}2} - p)] = [p(1-p)]^2/4$, since alleles

inherited from the father and from the mother are independent. Also, terms like $\mathrm{E}[(x_{\mathrm{M}1}-p)(x_{\mathrm{M}2}-p)(x_{\mathrm{F}1}-p)]=0$, since one factor is independent of the other two. Collecting together the various products gives (9.2).

9.2 IBD Probabilities at the Candidate Trait Locus

Given a a pair of affected sibs, let J_{M}, resp. J_{F}, be 1 or 0 according as the alleles at the trait locus from the mother, resp. from the father, are inherited IBD or not. Let $J=J_{\mathrm{M}}+J_{\mathrm{F}}$ denote the total number of alleles inherited IBD at a trait locus. Note that J_{M} and J_{F} are independent random variables taking values 0 and 1 with probability 1/2 each. The argument given above can be expressed conditionally, as $\mathrm{E}[(x_{\mathrm{M}1}-p)(x_{\mathrm{M}2}-p)|J_{\mathrm{M}}]=p(1-p)J_{\mathrm{M}}$. Other terms can be evaluated similarly, leading to $E(y_1y_2|J_{\mathrm{M}},J_{\mathrm{F}})=m^2+J\sigma_{\mathrm{A}}^2/2+J_{\mathrm{M}}J_{\mathrm{F}}\sigma_{\mathrm{D}}^2$, which in turn implies

$$\mathrm{E}(y_1y_2|J)=m^2+J\sigma_{\mathrm{A}}^2/2+I_{\{J=2\}}\sigma_{\mathrm{D}}^2\ , \tag{9.3}$$

where $I_{\{J=2\}}$ is the indicator of the event that the IBD count is two. Let $Q_2=\Pr(A)=\mathrm{E}(y_1y_2)$ be the probability given in (9.2) that both sibs are affected. Then by Bayes' formula:

$$\Pr(J=j|A)=\Pr(J=j)\frac{\Pr(A|J=j)}{\Pr(A)}=\Pr(J=j)\frac{E(y_1y_2|J=j)}{E(y_1y_2)}\ ,$$

Substituting (9.2) and (9.3), we find after some algebraic simplification that $\pi_j=\Pr(J=j|A)$ is given by

$$\begin{aligned}\pi_0&=[1-(\breve{\alpha}-\breve{\delta}/2)/Q_2]/4\ ,\\ \pi_1&=[1-\breve{\delta}/2Q_2]/2\ ,\\ \pi_2&=[1+(\breve{\alpha}+\breve{\delta}/2)/Q_2]/4\ ,\end{aligned} \tag{9.4}$$

where $\breve{\alpha}=(\sigma_{\mathrm{A}}^2+\sigma_{\mathrm{D}}^2)/2$ and $\breve{\delta}=\sigma_{\mathrm{D}}^2/2$. We have used the notation $\breve{\alpha}$, $\breve{\delta}$ because these quantities play roles in human genetics, here and in Chap. 11, similar to α, δ in the analysis of an intercross (cf. (9.6)). Note, however, that $0\le\breve{\delta}\le\breve{\alpha}$, although there is no similar restriction on α and δ. In the special case of an additive model, $\sigma_{\mathrm{D}}^2=0$, the equations simplify accordingly. While the terms in (9.4) are very simple, tedious calculation is required to evaluate them in terms of the allele frequency p and penetrances. Special cases are explored in the problems at the end of the chapter. A case of particular interest is the additive case, where $g_1=g_0+\alpha$, $g_2=g_0+2\alpha$, so $\sigma_{\mathrm{D}}^2=0$. Let $R_2=g_2/g_0=1+2\alpha/g_0$ denote the ratio of the penetrance of a DD-homozygote to that of a dd-homozygote. By solving for $\alpha=g_0(R_2-1)/2$, we find that $m=g_0+2p\alpha=g_0[1+p(R_2-1)]$ and $\sigma_{\mathrm{A}}^2=g_0^2p(1-p)(R_2-1)^2/2$. Thus $\sigma_{\mathrm{A}}^2/2Q_2$ and hence the IBD probabilities in (9.4) depend only on p and R_2

The case $R_2 = 1$, which is equivalent under the additive model to the case $g_0 = g_1 = g_2$, corresponds to no relation between the disease and the investigated gene. Indeed, when $R_2 = 1$, $\sigma_{\mathrm{A}}^2 = \sigma_{\mathrm{D}}^2 = 0$, so (9.4) gives the null distribution of the IBD status: $\pi_0 = \pi_2 = 1/4$, $\pi_1 = 1/2$. This distribution is the $B(2, 1/2)$ distribution. The expected number of alleles IBD in this case is 1. However, when $R_2 > 1$, the relation between the probabilities is $\pi_0 < \pi_2$. Also, since $\pi_1 = 1/2$, $\pi_2 = 1/2 - \pi_0$. The expected number of alleles IBD becomes $1/2 + 2\pi_2 = 1 + (\pi_2 - \pi_0) > 1$. Thus the equations (9.4) give quantitative meaning to the intuitive idea expressed in the introduction to this chapter that two siblings affected with the same disease are likely to have inherited the same disease predisposing allele from a parent.

The function "`DistIBD`" computes the IBD probabilities as a function of the allele frequency `p` and the penetrance probabilities `g0`, `g1` , and `g2`:

```
> DistIBD <- function(p,g0,g1,g2)
+ {
+     alpha <- (g2-g0)/2
+     delta <- g1 - g0 - alpha
+     m <- g0 + 2*p*alpha + 2*p*(1-p)*delta
+     a <- alpha + (1-2*p)*delta
+     d <- delta
+     sig.A <- 2*p*(1-p)*a^2
+     sig.D <- 4*p^2*(1-p)^2*d^2
+     Q <- m^2 + sig.A/2 + sig.D/4
+     pi.0 <- (1- (sig.A + sig.D/2)/(2*Q))/4
+     pi.1 <- (1-sig.D/(4*Q))/2
+     pi.2 <- (1 + (sig.A + 3*sig.D/2)/(2*Q))/4
+     return(data.frame(pi.0=pi.0,pi.1=pi.1,pi.2=pi.2))
+ }
```

Let us explore the effect of the parameters on the distribution of IBD for $g_0 = 0.05$ and $p = 0.1$, for an additive model with different values of α:

```
> alpha <- seq(0,0.4,by=0.1)
> IBD.prob <- DistIBD(0.1,0.05,0.05+alpha,0.05+2*alpha)
> IBD.e <- IBD.prob$pi.1+2*IBD.prob$pi.2
> IBD.sd <- sqrt(IBD.prob$pi.1+4*IBD.prob$pi.2 - IBD.e^2)
> round(cbind(alpha,IBD.prob,IBD.e,IBD.sd),3)
  alpha  pi.0 pi.1  pi.2 IBD.e IBD.sd
1   0.0 0.250  0.5 0.250 1.000  0.707
2   0.1 0.211  0.5 0.289 1.078  0.703
3   0.2 0.173  0.5 0.327 1.154  0.690
4   0.3 0.150  0.5 0.350 1.200  0.678
5   0.4 0.135  0.5 0.365 1.230  0.669
```

Of no surprise is the fact that the expectation increases with α. Note that the standard deviation remains more or less constant.

A more systematic exploration will also consider the effect of the allele frequency, which we vary from 0.1 to 0.5:

```
> alpha <- seq(0,0.4,by=0.02)
> g0 <- 0.05
> R2 <- 1+2*alpha/g0
> p <- seq(0.1,0.5,by=0.1)
> rep.a <- rep(alpha,length(p))
> rep.p <- rep(p,rep(length(alpha),length(p)))
> IBD.prob <- DistIBD(rep.p,g0,g0+rep.a,g0+2*rep.a)
> IBD.e <- IBD.prob$pi.1+2*IBD.prob$pi.2
> ncp <- (IBD.e-1)^2/0.5
> plot(range(R2),range(ncp),type="n",xlab="R2",ylab="ncp/n")
> for(i in 1:length(p)) lines(R2,ncp[rep.p==p[i]],lty=i)
> legend(1,max(ncp),legend=paste("p = ",p),lty=1:length(p))
```

The increase of the squared standardized difference of the expectation of IBD as a function of the parameter of genetic relative risk R_2 and for various values of allele frequency p is presented in Fig. 9.1. As we will see below, this parameter is the square of the noncentrality parameter of the test statistic and is the basis for the evaluation of the statistical power when scanning for disease predisposing genes. Roughly speaking, rarer disease susceptibility alleles with larger genetic relative risk are easier to detect.

9.3 A Test for Linkage at a Single Marker Based on a Normal Approximation

In the previous section we considered the distribution of a single quantity – the IBD status of a given pedigree – both under the null assumption $H_0 : \pi_2 = 1/4$, $\pi_1 = 1/2$, $\pi_0 = 1/4$ and under the additive alternative $H_1 : \pi_2 - \pi_0 > 0$. In this section we investigate the properties of a test statistic calculated from a sample of such observations. These properties form the basis for the justification of the use of the total number of alleles shared IBD as an appropriate test statistic under the additive model. We begin with the test for linkage of a single marker and move from there to consideration of tests using a genome scan.

Remark 9.1. It is not immediately clear that one can actually determine the number of alleles shared IBD. In fact, if one parent is homozygous at a marker, the sibs must inherit the same allele, and one cannot say with certainty whether it is inherited IBD. This problem makes for another level of difficulty in human genetics, compared to experimental genetics. For the moment we assume that this problem does not exist and return to it in Sect. 9.6.

In the case of siblings, the number of alleles IBD at a given marker locus can be 0, 1, or 2. Let N_0, N_1, and N_2 be the total number of pedigrees

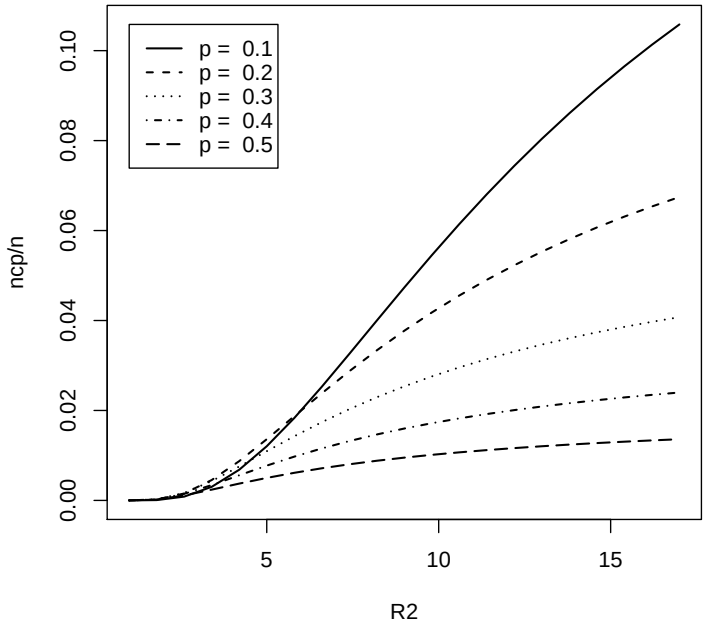

Fig. 9.1. The squared noncentrality parameter (per unit sample size) of the IBD statistic.

that share 0, 1, or 2 alleles IBD. The joint distribution of these counts is Multinomial(n, π), with $\pi = (\pi_0, \pi_1, \pi_2)$. Under the null distribution $\pi = (1/4, 1/2, 1/4)$. For an additive model (i.e., $\sigma^2_{\text{D}} = 0$) the alternative distribution at the trait locus takes the form $\pi = (1/4 - \sigma^2_{\text{A}}/(8Q), 1/2, 1/4 + \sigma^2_{\text{A}}/(8Q))$. The total IBD, standardized to have mean 0 and variance 1 under the hypothesis of no linkage, is a reasonable statistic for testing $H_0 : \sigma^2_{\text{A}} = 0$ versus the alternative $H_1 : \sigma^2_{\text{A}} > 0$.

The total number of alleles shared IBD is $N_1 + 2N_2$. Since the expected number of alleles shared IBD is n and the variance is $n/2$ (both computed under the null distribution), and since $n = N_0 + N_1 + N_2$, the standardized statistic is

$$Z = \frac{2N_2 + N_1 - n}{(n/2)^{1/2}} = \frac{N_2 - N_0}{(n/2)^{1/2}} \,. \tag{9.5}$$

Observe that the hypothesis tested is one sided, hence the null is rejected for large positive values of the test statistic. The threshold for significance is determined by the null distribution of the test statistic. As the sample size

increases ($n \to \infty$) the distribution of this test statistic resembles more and more that of the standard normal distribution.

The mean of the test statistic under the alternative hypothesis for a marker perfectly linked to the trait locus τ is given by

$$\xi = \mathrm{E}(Z_\tau) = (2n)^{1/2}(\pi_2 - \pi_0) = (n/2)^{1/2}\breve{\alpha}/Q_2 \; . \tag{9.6}$$

The variance of Z is $1 - (\breve{\alpha}/Q_2)^2/2 \approx 1$, for local alternatives. The noncentrality parameter at a linked marker is calculated below.

9.4 Genome Scans

For a genome scan, we use $\max_t Z_t$, where we now introduce the subscript t to denote marker location. The significance level and power can be found exactly as in Chaps. 4 and 6, provided we use an approximation suitable for a one-sided test and the appropriate value of β. It turns out that β is 0.04, exactly twice what it was for a backcross. The reason is that along each chromosome (one maternally inherited and the other paternally inherited), the two siblings involve two meiotic events. In contrast a backcross involved only one. A more detailed mathematical analysis follows.

To study the properties of Z_t and hence to approximate the significance level and power of a genome scan using the results of Chaps. 4 and 6, it is helpful to use the representation of the numerator $2N_{2t}+N_{1t}-n = \sum_{i=1}^{n}[J_i(t)-1]$. This representation shows that the correlation function and mean value of the standardized statistic Z_t can be obtained directly from the correlation function and mean value of each term in the numerator, namely $J(t)$, the number of alleles shared IBD by a sib pair at the marker t.

We first consider the case of markers that are unlinked to the trait locus. Under local alternatives, the same results for the covariance function hold at linked markers. Let $J(t)$ be written as $J_{\mathrm{M}}(t) + J_{\mathrm{F}}(t)$ where $J_{\mathrm{M}}(t)$ is the number of alleles, 0 or 1, inherited by the siblings IBD from their mother at locus t and $J_{\mathrm{F}}(t)$ is the number inherited from their father. Let s be a locus at recombination distance θ from t. If $J_{\mathrm{M}}(s) = 1$, then $J_{\mathrm{M}}(t) = 1$ if and only if both sibs have recombinations between t and s on their maternally inherited chromosome or neither sib does. The probability of no recombinations in the two maternal meioses is $(1-\theta)^2$, while the probability of two recombinations is θ^2. Thus $\Pr(J_{\mathrm{M}}(t) = 1|J_{\mathrm{M}}(s) = 1) = \theta^2 + (1-\theta)^2$. For future notational convenience, let φ be defined by $1-\varphi = \theta^2+(1-\theta)^2$. Similar reasoning applies to J_{F}, so $\Pr(J(t) = 2|J(s) = 2) = (1-\varphi)^2$. By similar arguments one sees that $\Pr(J(t) = 1|J(s) = 2) = 2\varphi(1-\varphi)$, $\Pr(J(t) = 1|J(s) = 1) = \varphi^2 + (1-\varphi)^2$, $\Pr(J(t) = 2|J(s) = 1) = \varphi(1-\varphi)$, etc., so we obtain a 3×3 matrix of transition probabilities from state $J(s) = i$ to state $J(t) = j$, for $i, j = 0, 1, 2$. Some calculation with these probabilities leads to

$$\mathrm{E}[J(t) - 1|J(s)] = (1 - 2\varphi)[J(s) - 1] \; . \tag{9.7}$$

Multiplying by $J(s) - 1$ and taking expectations, we obtain the important relation

$$\mathrm{cov}[J(t), J(s)] = (1 - 2\,\varphi)/2 = 2\,\theta(1 - \theta) = \exp(-0.04\,|s - t|)/2\ ,$$

where the third equality in the preceding expression follows from the equation $\theta = [1 - \exp(-0.02\,|t - s|)]/2$, for the recombination fraction θ in terms of genetic distance $|t - s|$ in cM. We conclude by observing that the preceding conditional probabilities and the resulting covariance are exactly the same as for $x(t)$, the number of A alleles in an intercross design, except that θ has been replaced by φ, which has the effect of turning the parameter 0.02 into 0.04 in the exponent of the correlation coefficient.

As an illustration, let us determine the thresholds for a genome scan with various inter-marker spacings. Note that the genetic length of the human genome is very roughly about twice that of a mouse, or about 3,200 cM. Moreover, the genetic material in humans is distributed among 23 pairs of chromosomes (22 pairs of autosomes and a pair of sex chromosomes). We use these values in the approximation (**??**), but we divide the expression in the exponent by 2, since we are now interested in a one-sided test, and we set $\beta = 0.04$, to obtain:

```
> Delta <- c(35,20,10,5,1)
> z <- vector(length=length(Delta))
> names(z) <- paste("Delta=",Delta,sep="")
> for (i in 1:length(Delta)) z[i] <-
+     uniroot(OU.approx,c(3,4),beta=0.04,Delta=Delta[i],
+     length=3200,chr=23,center=0.05,test="one-sided")$root
> round(z,3)
Delta=35 Delta=20 Delta=10  Delta=5  Delta=1
   3.337    3.459    3.601    3.721    3.906
```

The noncentrality parameter for a marker at no recombination distance from the trait locus itself was given in the preceding section. To evaluate power in a genomic scan, we must also know the effect of recombination on the noncentrality parameter. Observe that the reasoning behind (9.7), which depends only on the recombination fraction between the loci s and t continues to apply if we set $s = \tau$, the trait locus. By taking expectations, we conclude that at a marker t linked to the trait locus τ,

$$\mathrm{E}(Z_t) = \mathrm{E}(Z_\tau)(1 - 2\varphi) = \xi \exp(-0.04|t - \tau|)\ .$$

The rate of decay of the noncentrality parameter is twice what it was for a backcross or an intercross. This means that as one increases the inter-marker distance, there is a greater loss of power for detecting a trait locus midway between markers in sib pairs than in a backcross or an intercross.

We now explore numerically the power to detect a trait locus in a genome scan as a function of the noncentrality parameter at the trait locus. We consider separately the case where the trait is perfectly linked to a marker and

the case where the trait locus is midway between two markers. The functions "power.marker" and "power.midway", which were developed in Chap. 6, are used in order to obtain analytic approximations:

```
> xi <- seq(3,6,by=0.1)
> power.mark <- matrix(nrow=length(xi),ncol=length(Delta))
> colnames(power.mark) <- names(z)
> power.between <- power.mark
> for (j in 1:length(Delta))
+ {
+     power.mark [,j] <- power.marker(z[j],0.04,Delta[j],xi)
+     for (i in 1:length(xi)) power.between[i,j] <-
+         power.midway(z[j],0.04,Delta[j],xi[i])
+ }
```

Let us plot the power functions for the two cases:

```
> old.par = par(mfrow=c(1,2))
> plot(range(xi),c(0,1),type="n",xlab="xi",ylab="power",
+     main="Trait locus on a marker")
> for(j in 1:length(Delta)) lines(xi,power.mark[,j],lty=j)
> legend(3.5,0.3,bty="n",legend=names(z),lty=1:length(z))
> plot(range(xi),c(0,1),type="n",xlab="xi",ylab="power",
+     main="Trait locus between markers")
> for(j in 1:length(Delta)) lines(xi,power.between[,j],lty=j)
> par(old.par)
```

The output is given in Fig. 9.2. Observe the relatively small decrease in power as the markers density increases in the case that the trait locus is perfectly linked to a marker compared to a more substantial increase in the power when the locus is between markers. The difference is more pronounced than for a backcross or intercross. The reason is the greater recombination parameter ($\beta = 0.04$ instead of 0.02) in sib pairs, because two parental meioses are involved. The result is a more rapid decay of the linkage signal as one moves away from a marker, so that markers in sib pairs about 10 cM apart lead to about the same loss of power to detect a gene at the midpoint as markers about 20 cM apart in a backcross. This difference is even larger when we consider in the following section markers that are only partly informative. As a result, for a genome scan based on affected sib-pairs, it appears advisable to use a substantially denser collection of markers than we found necessary for a backcross or intercross in experimental genetics. The results of the following section will reinforce this conclusion.

An interesting example is provided by John *et al.* [43], who reanalyzed with SNPs an earlier study of rheumatoid arthritis that had been conducted with microsatellite markers at an approximately 10 cM inter-marker spacing. The earlier analysis gave the value $\max_t Z_t = 4.22$ on chromosome 6 (in the Human Leucocyte Antigen, or HLA, region, which harbors a large number of

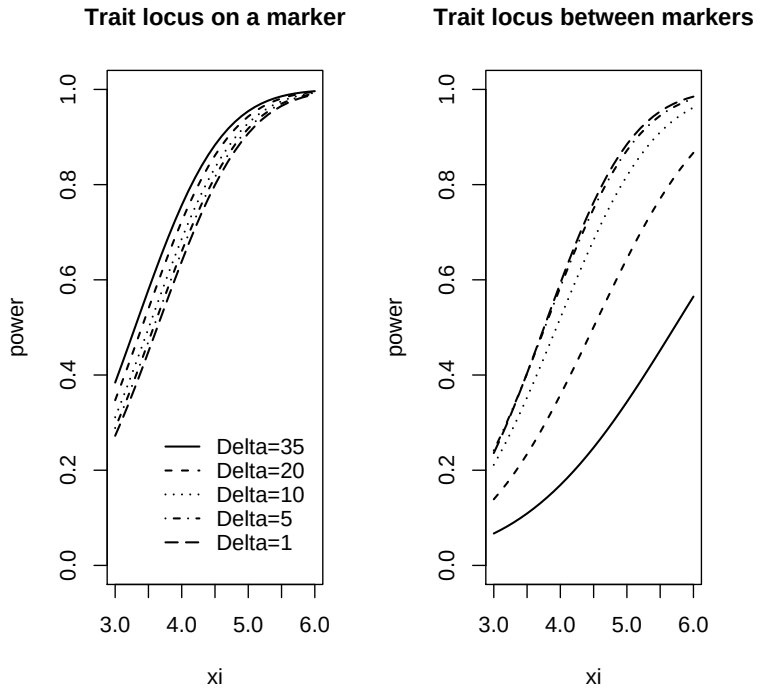

Fig. 9.2. The power for detecting a trait locus in a genome scan

genes associated with the function and diseases of the immune system). The p-value based on the approximation in Chap. 4 would be 0.004.

SNPs at an inter-marker distance of about 0.3 cM led to $\max_t Z_t = 3.97$ (at very close to the same genomic location), with a p-value of just slightly more than 0.05. Since markers placed this close together might fail to satisfy the important assumption of linkage equilibrium, the authors selected one (particularly informative) marker per cM. This reduced the value of $\max_t Z_t$ to 3.54, which has a p-value of 0.18.

Although these results suggest that the SNPs overall were not quite as efficient as SSRs, the authors indicated an overall preference for SNPs. A statistical reason for this preference is that the profile of the process Z_t seemed to give a more precise picture of the location of the gene with the more closely spaced markers. A scientific reason is that the slight discrepancy between the locations of the peaks for the 10 cM and the 0.3 cM scans seemed to suggest that the latter provided a better estimate of the location of the appropriate gene, which could be regarded as known from earlier studies.

9.5 Parametric Methods

The approach described above is frequently called *nonparametric* to distinguish it from the *parametric* approach pioneered by Morton [56] and systematically described by Ott [57]. While parameters appear in both approaches, in the parametric approach the latent *genetic* parameters – penetrances and allele frequencies – play an explicit role in statistics to detect linkage, whereas in the nonparametric approach one concentrates on *statistical* parameters, which can in principle be estimated directly from experimental data without a specific genetic model. Examples are the frequency of a trait, or the probability that a particular relative of an affected is also affected, which can be estimated from population samples of phenotypes, or noncentrality parameters of statistics to detect linkage, which can be estimated from a combination of phenotypic and genotypic data. The parametric approach was developed at a time when the diseases under consideration involved a single gene, often showed a clear mode of inheritance, dominant or recessive, and had essentially complete penetrance, so it might be reasonable to assume $g_0 \approx 0$, with $g_1 = g_2 \approx 1$ for a dominant trait or $g_1 \approx 0, g_2 \approx 1$ for a recessive trait. Moreover, the number of markers available was extremely small, so the limiting factor in one's ability to map a disease gene was not the genetic effect on the trait, which was quite pronounced, but the recombination distance of the nearest marker to the gene.

For example, suppose we see the disease occurring in successive generations of pedigrees with approximately 50% of the offspring of affected individuals also having the disease. This suggests that the trait is dominant and fully penetrate without sporadic cases ($g_0 = 0$), and under an assumption of Hardy-Weinberg equilibrium we can estimate the frequency p of the disease gene from the population prevalence, $2p(1-p) + p^2 = 2p - p^2$, of the trait.

The simplest illustration of a parametric analysis arises from considering a three generation pedigree of an affected grandparent, say the grandfather, the intervening parent, say the father, and an affected grandchild. We assume the grandmother and mother are unaffected, so under the assumption of full penetrance the father is also affected. Assume there is a marker that has recombination θ with the trait, and assume that all marker alleles in the pedigree are unique, so we know exactly which marker allele passes from the affected grandfather, to the father, and the allele that the father then passes to the grandchild. If in addition to the disease allele, the father passes to his child the marker allele he received from the grandfather, then by definition there is no recombination between the disease allele and the marker allele. This happens with probability $1 - \theta$. If the father passes the allele he inherited from the grandmother, there is recombination with the disease allele, which happens with probability θ. If we let R denote the number of recombinations, 0 or 1, we have a Bernoulli variable with $\Pr(R = r) = \theta^r(1-\theta)^{1-r}$. If we have a sample of n such grandchildren, the total number of recombinations in the sample would be binomial with probability θ, and we could test the hypoth-

esis $\theta = 1/2$ by counting these recombinations. Observe that the grandfather and grandson share one allele IBD at the marker locus if and only if $R = 0$. Hence a test based on the number of recombinations is equivalently a test based on the number of alleles shared IBD, so the parametric analysis is in this case equivalent to a nonparametric analysis based on number of alleles shared IBD. Observe also that in this scenario there can be multiple affected grandchildren in a pedigree without any essential changes, and that unaffected grandchildren also provide linkage information. For them, however, recombinations have probability $1 - \theta$ and non-recombinations have probability θ, since they are assumed not to have inherited the disease allele. Finally note that this simple scenario would become substantially more complex if there are sporadic cases and/or the penetrance of the disease allele is less than one, since then we could not be sure that the grandchild has the disease allele nor that it comes from the grandfather. In that case an appropriate parametric likelihood function would involve the penetrances g_0, g_1, g_2 and the allele frequency p in addition to θ.

Returning to the case of affected sib pairs, to put parametric and nonparametric analysis on similar footing we assume a two generation pedigree, but parental phenotypes are unknown. If there are few sporadic cases in the population and the disease is rare (and dominant), we would be willing to suppose that there is exactly one copy of the disease allele in the parental generation. But we do not know which of four marker alleles present in the parents and lying on the same chromosome as the disease related locus (we assume as above that markers are completely informative) is actually linked to the disease allele itself. (Genetic terminology is that the *linkage phase* is unknown.) Hence we consider the four possibilities to be equally likely. The analysis is more complicated but similar in principle to that given above. It leads to the number of alleles shared IBD by the siblings, so we eventually arrive at the probabilities in (9.4), or more generally the corresponding probabilities, say $\pi_i(\theta)$ for a marker at recombination distance θ from the disease gene. These probabilities would be assumed known except for the parameter θ. To test the null hypothesis $\theta = 1/2$ (no linkage) against a specific alternative value $\theta < 1/2$, we could use the log likelihood ratio statistic

$$\sum_{j=0}^{2} N_j \log[\pi_j(\theta)/\pi_j(1/2)] \,, \tag{9.8}$$

which is often maximized with respect to θ to reflect the fact that the true θ is almost always unknown. Under the conditions described above, the noncentrality parameter of this statistic is typically large if θ is small, but would be small if the true θ is close to $1/2$.

A detailed development of this approach, especially the modifications required to deal with the fact that for complex diseases the allele frequencies and penetrances are essentially never known, is beyond the scope of this book. The most important strength of a parametric approach, to which we return in

Chap. 11, is that, subject to being able to perform the required calculations, it generalizes directly to pedigrees having varying numbers and configurations of affecteds and to arbitrary combinations of pedigrees. This property has played an important role in dealing with single gene traits of large penetrance, where large pedigrees with multiple affecteds are common. It is less important for dealing with complex diseases, involving small penetrances where pedigrees with large numbers of affecteds are rare.

The weakness of a parametric approach is that for complex diseases there may be multiple genes of incomplete penetrance that may interact with each other or with the environment, as well as non-genetic cases of the disease ($g_0 > 0$), with the result that one has no clear idea of the number or values of the relevant penetrances and allele frequencies. In addition, since modern DNA analysis has made available a large number of mapped markers, the emphasis on testing an hypothesis about θ seems misplaced. We are prepared to assume that *some* markers are closely linked to the relevant genes. However, the complexity of the genetics can lead to small noncentrality parameters, even at tightly linked markers, so the true signal at a linked marker may be small compared to the apparent signals arising from chance fluctuations at spurious markers throughout the genome. Hence in our outlook we have emphasized a null hypothesis to the effect that at the marker locus under consideration there is effectively no departure from Mendelian segregation of genotypes, so the noncentrality parameter of any test statistic is zero.

To gain somewhat more insight into the nature of a parametric analysis and prepare for a related discussion in Chap. 11, suppose (to simplify calculations) that $\check{\delta} = 0$ and $\check{\alpha}$ is small. By using the Taylor series approximation $\log(1 + x) \approx x - x^2/2$, valid for small $|x|$, one can show that the log likelihood ratio statistic at a marker locus t assumed to be a recombination distance of θ from the trait locus is approximately

$$\xi(1 - 2\varphi)Z_t - \xi^2(1 - 2\varphi)^2/2 \tag{9.9}$$

where Z_t is the approximately normal statistic defined above and $\xi = (n/2)^{1/2}(\check{\alpha}/Q_2)$ is its noncentrality at the trait locus τ. So far we have regarded ξ as known. If we admit that it is unknown, the parameters ξ and φ cannot be estimated separately if we only observe Z_t. Only the parameter $\eta = \xi(1 - 2\varphi)$ can be estimated.

At this point there are different possibilities for proceeding. (i) If we maximize the preceding expression with respect to $\eta = \xi(1 - 2\varphi) \geq 0$, we get $[\max(Z_t, 0)]^2/2$, where the nonnegativity restriction arises from the fact that the parameter η cannot be negative. This would be equivalent to the statistic Z_t (for one-sided alternatives), but as we shall see, the equivalence for this simple problem turns out to be the exception, not the rule. (See Prob. 9.11 for an example and the related discussion in Chap. 11.) (ii) If we take the attitude that markers are reasonably dense, so the distance from the nearest marker to the trait locus is likely to be small, we might simply set $\varphi = 0$ in (9.9) and scan the genome for maxima with respect to t. The result is a monotonic

function of $\max Z_t$, hence is again equivalent to using $\max Z_t$ directly. (iii) If we take into consideration that we have a collection of mapped markers, the situation is similar to that in Chap. 7. For the asymptotic Gaussian model, where the log likelihood at τ is given by (9.9) with $t = \tau$ and $\varphi = 0$, we can also compute the likelihood function for a trait locus τ lying between markers t_i and t_{i+1}. Using the notation of Chap. 7 (but with φ in place of θ), we have that $\mathrm{E}_0[Z_\tau | Z_{t_i}, Z_{t_{i+1}}] = \sigma_\tau' W Z$, $\mathrm{var}_0[Z_\tau | Z_{t_i}, Z_{t_{i+1}}] = 1 - \sigma_\tau' W \sigma_\tau$, and hence the likelihood function equals

$$\mathrm{E}_0[\exp(\xi Z_\tau - \xi^2/2) | Z_{t_i}, Z_{t_{i+1}}] = \exp(\xi \sigma_\tau' W Z - \xi^2 \sigma_\tau' W \sigma_\tau / 2) \; .$$

For ξ regarded as known, the likelihood ratio statistic for a genome scan would be the maximum over τ of the expression appearing in the exponent. If ξ is regarded as unknown, we can maximize over $\xi \geq 0$ as well. This leads to a one-sided version of the statistic in Chap. 7. Based on the results of Chap. 7, it seems unlikely that these statistics will be substantially more powerful than the simpler statistics studied earlier in this chapter.

9.6 Estimating the Number of Alleles Shared IBD

In general, IBD status, which is the basis for the statistics discussed above, is not observed directly. It needs to be inferred from genotype information. Assume that both parents of the siblings were recruited and the genotypes of all the given members of the family were obtained. If multi-allelic markers are used, one may be able to observe that at a particular locus the two parents have four distinct alleles. This favorable scenario enables the precise determination of the IBD status of the siblings. At the other extreme, if both parents are homozygous, then the specific marker provides no information regarding the IBD status of the siblings at that locus. The marker is then said to be *uninformative*. There may also exist intermediate cases, e.g., where one parent is homozygous while the other is heterozygous, or when both parents are heterozygous for the same two alleles. Such markers are denoted *partially informative*. However, even when a marker is partially informative or totaly uninformative, there may be other markers nearby which are either informative or at least partially informative. If these markers are sufficiently close, so there is little chance of recombination, one may attempt to infer the IBD status at the given locus based on the genotypes at those nearby loci and then conduct a genome scan with reconstructed IBD statistics. The problem of partial information regarding IBD relations among the affected siblings and the need to exploit genotype information from nearby markers in order to reconstruct the IBD becomes even more acute if the parents are not available for genotyping. This is commonly the case in late onset diseases, such as Alzheimer disease, in which the participating affected siblings are typically older and are less likely to have living parents.

Recall that markers fall into two main classes: SNPs, which are bi-allelic, and various classes of multi-allelic markers, e.g., SSRs, which often have 4–10 alleles. While SNPs are much more numerous and more easily genotyped, they are individually less informative. In most of the following we concentrate on SNPs and find that because each one by itself is relatively uninformative, there is a considerable loss of information unless they are reasonably dense. The programs are easily adapted to multi-allelic markers and show that SSRs can be more widely separated without a corresponding loss of information.

In this section we will investigate the effect of partial information regarding the IBD relations on the statistical properties of the test statistics. We will substitute for the unknown IBD statistic its conditional expectation, given the genotypic information at hand. This is similar to the case of missing genotypes that was discussed in Chap. 7. However, the computation of the conditional expectation is more complex and will require application of algorithms that were originally developed in the context of what are called *hidden Markov models*, or HMM in short.

The section is divided into three subsections. In the first subsection we will develop `R` code for the simulation of pedigrees. The HMM algorithms will be presented in the second subsection. In the third subsection we will explore the statistical properties of a genome scan with affected sib pairs when only their genotypes are available. The tools that were developed in the first two subsections will be used in that exploration.

9.6.1 Simulating Pedigrees

Our first goal is to develop `R` code that will enable us to simulate affected sib pairs. The programs we develop are similar to those developed in the chapters that dealt with experimental genetics. However, there is a major difference between the situation we previously considered and the current one. In the experimental designs that we considered before, subjects were not preselected based on their phenotypes. In particular, the segregation of genetic material from one generation to the next followed Mendel's segregation rules. In the case at hand, however, the subjects are selected because they express the trait (a disease). This selection rule results in a distortion of the segregation of alleles in loci linked to trait-related genes. In fact, it is exactly this distortion that allows us to detect such loci. As a result, we now need to rewrite our programs in order to allow for distortion in the segregation in the presence of a trait locus.

Start with an adaptation to the new setting of the function "`meiosis.chr`", which simulates the gamete being segregated from a parent to an offspring:

```
> meiosis.link <- function(GF,GM,markers,qtl,inhe)
+ {
+     n <- nrow(GF)
+     GS <- GF
```

```
+     loci <- sort(c(qtl, markers))
+     rec.frac <- (1-exp(-0.02*diff(loci)))/2
+     index <- 1:length(markers)
+     from.GM <- inhe
+     for (i in index[markers >= qtl])
+     {
+         rec <- rbinom(n,1,rec.frac[i])
+         from.GM <- from.GM*(1-rec) + (1-from.GM)*rec
+         GS[from.GM==1,i] <- GM[from.GM==1,i]
+     }
+     from.GM <- inhe
+     for (i in rev(index[markers < qtl]))
+     {
+         rec <- rbinom(n,1,rec.frac[i])
+         from.GM <- from.GM*(1-rec) + (1-from.GM)*rec
+         GS[from.GM==1,i] <- GM[from.GM==1,i]
+     }
+     return(GS)
+ }
```

Observe that in the default application in the definition of the function "`mating`" below, which corresponds to the null case of no trait locus and which obeys Mendel's laws of segregation, the inheritance vector at the first marker consists of realizations of independent 0-1 random variables. The segregation of the rest of the markers is determined in the first loop according to the process of recombination in exactly the same way it was done in the function "`meiosis.chr`". The second loop is not activated.

If a trait related locus does exist at a locus denoted "`qtl`", then the selection rule may distort the distribution of the inheritance indicator at that locus. This distorted distribution will be generated externally, and the resulting vector of the inheritance indicator at the trait locus may be imported in the argument "`inhe`". The distortion is reflected at the markers on both sides of the trait locus due to linkage and the process of recombination. The process to the right of the trait locus is generated in the first loop and the process to the left is generated in the second loop.

Subjects' pairs of parental gametes are stored in a list. This list contains two matrices, one for each gamete. The columns of the matrices correspond to the markers and the rows to independent copies. The function "`mating`" is an adaptation to the new setting of the function "`cross`" which was applied in the context of experimental genetics. It takes as input two subjects (a father and a mother) and returns as output a new subject (an offspring):

```
> mating <- function(fa,mo,markers,qtl=markers[1],
+     inhe.fa=rbinom(nrow(fa$pat),1,0.5),
+     inhe.mo=rbinom(nrow(mo$pat),1,0.5))
+ {
```

```
+     pat <- meiosis.link(fa$pat,fa$mat,markers,qtl,inhe.fa)
+     mat <- meiosis.link(mo$pat,mo$mat,markers,qtl,inhe.mo)
+     return(list(pat=pat, mat=mat))
+ }
```

As an illustration of the application of the code let us generate the processes of IBD on a given chromosome for 10 independent pedigrees with markers spaced 20 cM apart:

```
> n.ped <- 10
> markers <- seq(0,140,by=20)
> n.mark <- length(markers)
> fa <- list(pat=matrix(1,n.ped,n.mark),
+     mat=matrix(2,n.ped,n.mark))
> mo <- list(pat=matrix(3,n.ped,n.mark),
+     mat=matrix(4,n.ped,n.mark))
> sib1 <- mating(fa,mo,markers)
> sib2 <- mating(fa,mo,markers)
> ibd <- (sib1$pat==sib2$pat)+
+     (sib1$mat==sib2$mat)
> ibd
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
 [1,]    0    1    0    0    0    1    1    1
 [2,]    2    2    1    0    0    0    1    1
 [3,]    1    1    0    0    0    1    1    1
 [4,]    2    2    1    2    0    0    2    2
 [5,]    2    2    2    1    1    1    1    1
 [6,]    1    1    1    1    2    1    0    1
 [7,]    0    0    0    1    1    1    2    2
 [8,]    1    1    2    1    1    1    1    2
 [9,]    1    2    2    2    2    1    1    0
[10,]    1    1    2    1    0    0    0    0
```

Here the markers are fully informative and the IBD process can be computed directly from the alleles of the offspring.

Next we turn to the simulation of pedigrees under the alternative distribution. In Sect. 9.2 we investigated the distribution of the IBD status at the trait locus as a function of the allele and penetrance frequencies, given that both siblings are affected. The distribution of the inheritance vectors is a reflection of the IBD distribution. The inheritance vector has four components, indicating the parental source of the gamete segregated from (i) the father and (ii) the mother of the first sibling and (iii) the father and (iv) the mother of the second sibling. Observe that the marginal probability of each of the grandpaternal origins in components (i)–(iv) is 1/2. The IBD relation introduces dependence between the components. When the IBD status equals zero, then the parental sources for (i) and (iii) and for (ii) and (iv) are opposite. When the IBD status equals two, the parental sources are the same. When

the IBD status equals one, then for one pair the parental source is the same and for the other it is opposite:

```
> inhe.vector <- function(ibd.prob,n.ped)
+ {
+    ibd.qtl <- sample(0:2,n.ped,replace=TRUE,prob=ibd.prob)
+    sib1.pat <- rbinom(n.ped,1,0.5)
+    sib1.mat <- rbinom(n.ped,1,0.5)
+    pat.equal <- rbinom(n.ped,1,0.5)
+    sib2.pat <- sib1.pat*pat.equal+(1-sib1.pat)*(1-pat.equal)
+    sib2.mat <- sib1.mat*(1-pat.equal)+(1-sib1.mat)*pat.equal
+    inhe <- cbind(sib1.pat,sib1.mat,sib2.pat,sib2.mat)
+    inhe[ibd.qtl==0,3:4] <- 1-inhe[ibd.qtl==0,1:2]
+    inhe[ibd.qtl==2,3:4] <- inhe[ibd.qtl==2,1:2]
+    return(inhe)
+ }
```

The program works by simulating first the independent components (i) and (ii) from the marginal Bernoulli distribution. Using random assignment, some of the rows are set to have (iii) match (i) and (iv) opposite to (ii) while the other rows are set in the opposite way. This corresponds to having IBD equal to one. Subsequently, the relations between the components are changed to the appropriate relations in the rows where IBD is equal to zero and in the rows where it is equal to two. The IBD itself is generated from the appropriate alternative distribution, which is provided in the argument "`ibd.prob`". To illustrate consider the additive model in which we take $p = 0.2$, $g_0 = 0.1$, $\alpha = 0.4$:

```
> n.ped <- 10^5
> ibd.prob <- DistIBD(0.2,0.1,0.5,0.9)
> qtl <- 80
> inhe.qtl <- inhe.vector(ibd.prob,n.ped)
> fa <- list(pat=matrix(1,n.ped,n.mark),
+             mat=matrix(2,n.ped,n.mark))
> mo <- list(pat=matrix(3,n.ped,n.mark),
+             mat=matrix(4,n.ped,n.mark))
> sib1 <- mating(fa, mo, markers,
+             inhe.fa=inhe.qtl[,"sib1.pat"],
+             inhe.mo=inhe.qtl[,"sib1.mat"],qtl=qtl)
> sib2 <- mating(fa, mo, markers,
+             inhe.fa=inhe.qtl[,"sib2.pat"],
+             inhe.mo=inhe.qtl[,"sib2.mat"],qtl=qtl)
> ibd <- (sib1$pat==sib2$pat)+(sib1$mat==sib2$mat)
> sum(ibd.prob*0:2)
[1] 1.137339
> 2*mean(inhe.qtl[,1:2]==inhe.qtl[,3:4])
```

```
[1] 1.1369
> round(apply(ibd,2,mean),4)
[1] 0.9996 1.0098 1.0261 1.0622 1.1369 1.0605 1.0288 1.0108
```

A susceptibility locus is present 80 cM from the telomere, next to the 5th marker. Note that the average IBD at the marker is about equal to the expectation computed from the IBD probabilities. The expected IBD is elevated in the vicinity of the trait locus and it gradually decreases to the null expectation as markers become more distant from that locus.

Now consider the replacement of the fully informative markers by partially informative ones. The information provided by markers is in the form of the classification of pedigrees based on the genotypes of those members for which the genotypes are obtained. For example, we will assume in the sequel that genotypes are obtained for both siblings but not for their parents. We will also assume that the markers have `n.al` distinct alleles, with the default value of two. Genotype measurement for an individual returns the combined reading of its pair of homologous chromosomes, without distinguishing the parental source. Hence, for bi-allelic markers one may obtain three distinct genotypes. More generally, for markers with `n.al` alleles the total number of distinct genotypes is $\texttt{n.al}(\texttt{n.al}+1)/2$. The total number of genotypes for pair of siblings is the square of the number of individual genotypes.

We will find it easier to simulate and compute the distribution of the four parental alleles of the two siblings. However, it should be realized that these alleles are not observable. Instead, what one gets to observe are the genotypes, which are a many-to-one mapping of the four alleles. As a first step we introduce a function that maps alleles to genotypes:

```
> genotype <- function(a1,a2,a3,a4,n.al=2)
+ {
+     a.m <- pmin(a1,a2)
+     a.M <- pmax(a1,a2)
+     g1 <- a.M + (a.m-1)*(n.al-a.m/2)
+     a.m <- pmin(a3,a4)
+     a.M <- pmax(a3,a4)
+     g2 <- a.M + (a.m-1)*(n.al-a.m/2)
+     g <- g1 + (g2-1)*n.al*(n.al+1)/2
+     return(g)
+ }
```

In the specific case of a bi-allelic markers the function returns the combined genotypes coded as an integer between one and $3^2 = 9$.

The function "`ped.geno`" takes as input a pair of siblings and a vector of population allele distribution. It produces as an output a matrix of coded genotypes. Each column of the matrix corresponds to a marker and each row corresponds to a pedigree. Markers are assumed to be identically distributed and in linkage equilibrium, and pedigrees are assumed to be unrelated (which means statistical independence):

```
> ped.geno <- function(sib1,sib2,f=rep(1/2,2))
+ {
+     n.ped <- nrow(sib1$pat)
+     n.mark <- ncol(sib1$pat)
+     n.al <- length(f)
+     par.al <- list()
+     for(par in 1:4) par.al[[par]] <-
+         matrix(sample(1:n.al,n.ped*n.mark,
+             replace=TRUE,prob=f),n.ped,n.mark)
+     a <- inhe <- c(sib1,sib2)
+     for (v in 1:4) for (par in 1:4)
+         a[[v]][inhe[[v]]==par] <-
+             par.al[[par]][inhe[[v]]==par]
+     geno <- genotype(a[[1]],a[[2]],a[[3]],a[[4]],n.al)
+     return(geno)
+ }
```

The function works by simulating alleles (integers in the range between 1 and `n.al`) for each of the four parental gametes. The function "`sample`" is used in order simulate the alleles from the population distribution of marker alleles. The lists "`sib1`" and "`sib2`" store two matrices each with the index of the parental source of the gamete. The resulting inherited alleles are computed and sorted in the list "`a`". Finally, the function "`genotype`" is applied in order to compute the resulting genotype codes.

9.6.2 Computing the Conditional Distribution of IBD

The goal in this subsection is to reconstruct the unobserved process of IBD in a pedigree using the marker genotypes. This will be conducted by the calculation of the conditional expectation of the full-information statistic – the total number of alleles inherited IBD for the two siblings – given the genotypic information at hand. This conditional expectation is straightforward to compute once the conditional distribution of IBD, given the genotypes, is known. The calculation of the latter is the subject of this subsection.

The probabilistic structure of the observations may be modeled using a "hidden" process. The hidden process is the process of IBD at the markers. This process may not be observed directly, but it does have an effect on the distribution of the observed genotypes. This effect may be exploited in order to make inference on the underlying hidden process. In particular, when the underlying hidden process is Markovian and the distribution of an observation is determined by the state of the hidden process at the location of the observation, independently of its values at other locations, the model is called a hidden Markov model (HMM). Our case fits into this setting since the process of IBD is Markovian and since markers were assumed to be in linkage equilibrium.

Table 9.1. Conditional distribution of genotypes of ASP, given the IBD status.

	IBD=0	IBD=1	IBD=2
1=(0,0)	$(1-f)^4$	$(1-f)^3$	$(1-f)^2$
5=(1,1)	$4f^2(1-f)^2$	$f(1-f)$	$2f(1-f)$
9=(2,2)	f^4	f^3	f^2
2=(0,1)	$2f(1-f)^3$	$f(1-f)^2$	0
4=(1,0)	$2f(1-f)^3$	$f(1-f)^2$	0
6=(1,2)	$2f^3(1-f)$	$f^2(1-f)$	0
8=(2,1)	$2f^3(1-f)$	$f^2(1-f)$	0
3=(0,2)	$f^2(1-f)^2$	0	0
7=(2,0)	$f^2(1-f)^2$	0	0

The distribution of a HMM is fully determined by the initial distribution and the transition matrices of the underlying Markov process and by the conditional distribution of the observations, given the states of the underlying process. The latter refers in our case to the conditional distribution of the pair genotypes of the siblings, given the IBD status at the marker.

The conditional distribution of the genotypes of the siblings, given the IBD process, is a function of the frequency of the alleles in the population. If there is no identity-by-descent among the alleles of the siblings (IBD=0), the two genotypes are independent and follow the Hardy-Weinberg distribution. In the case where exactly one pair of alleles has a common source (IBD=1), then the other two alleles (one in each sibling) are independent of each other and of the IBD allele. Finally, when each of the two alleles in one sibling has a matching IBD allele in the other sibling (IBD=2), then the genotypes of the two siblings fully match. The distributions of the siblings' genotypes for a bi-allelic marker and for each of the IBD situations is given in Table 9.1. To avoid confusion we have used the letter f to represent the population frequency of the allele of the marker. In contrast, we have used the letter p to represent the frequency of the allele D of the trait locus.

The function "`geno.given.ibd`" computes this table for allele frequencies denoted by the vector `f` (with the default of uniformly distributed bi-allelic marker):

```
> geno.given.ibd <- function(f=c(0.5,0.5))
+ {
+    n.al <- length(f)
+    P.0 <- outer(outer(f,f),outer(f,f))
+    P.1 <- P.2 <- array(0,dim=rep(n.al,4))
+    for(a2 in 1:n.al) for(a1 in 1:n.al)
+    for(a3 in 1:n.al) for(a4 in 1:n.al)
+    {
+       if (a1==a3 & a2==a4)
+       {
+          P.2[a1,a2,a3,a4] <- f[a1]*f[a2]
```

```
+          P.1[a1,a2,a3,a4] <- f[a1]*f[a2]*(f[a1]+f[a2])/2
+       }
+       if (a1==a3 & a2!=a4)
+       {
+          P.1[a1,a2,a3,a4] <- f[a1]*f[a2]*f[a4]/2
+       }
+       if (a1!=a3 & a2==a4)
+       {
+          P.1[a1,a2,a3,a4] <- f[a1]*f[a3]*f[a2]/2
+       }
+    }
+    a1 <- rep(1:n.al,n.al^3)
+    a2 <- rep(rep(1:n.al,rep(n.al,n.al)),n.al^2)
+    a3 <- rep(rep(1:n.al,rep(n.al^2,n.al)),n.al)
+    a4 <- rep(1:n.al,rep(n.al^3,n.al))
+    geno <- genotype(a1,a2,a3,a4,n.al)
+    P.0 <- tapply(as.vector(P.0),geno,sum)
+    P.1 <- tapply(as.vector(P.1),geno,sum)
+    P.2 <- tapply(as.vector(P.2),geno,sum)
+    P <- cbind(P.0,P.1,P.2)
+    colnames(P) <- paste("State=",0:2,sep="")
+    return(P)
+ }
```

The function works by computing the distribution of the vector of four alleles for the two siblings in each of the IBD situations. The results are stored in three vectors, each of length `n.al` to the fourth power. The probabilities for the different genotypes are computed by summation of the four allelic probabilities according to the level of the genotype. This is carried out with the aid of the function "`tapply`". The function "`tapply`" takes as input a vector, a factor, and a function. The function is applied to the collection of values of the vector that correspond to each given level of the factor.

The other component is the distribution of the unobserved IBD process. This process is generated in the case under consideration as a function of the four inheritance indicators. Each of these indicators can be viewed as an independent process with two states (0 or 1). The states of the processes may change from one marker to the next, depending on the recombination fraction. Under a model of no crossover interference (the Haldane model) these independent inheritance processes are Markovian. As it turns out, for the specific pedigree structure of two siblings the process of IBD is Markovian as well. The transition matrix of going from one marker to the next, which was discussed on our analysis of J in Sect. 9.3, is computed in the function "`trans.mat`":

```
> trans.mat <- function(theta)
+ {
+     phi <- 1-theta^2-(1-theta)^2
+     Tr <- matrix(c((1-phi)^2,2*phi*(1-phi),phi^2,
+            phi*(1-phi),phi^2+(1-phi)^2,phi*(1-phi),
+            phi^2,2*phi*(1-phi),(1-phi)^2),3,3,byrow=TRUE)
+     colnames(Tr) <- paste("to.IBD=",0:2,sep="")
+     rownames(Tr) <- paste("from.IBD=",0:2,sep="")
+     return(Tr)
+ }
```

For example, when the distance between two markers is 20 cM, then the fraction of recombination and the transition matrix are given by:

```
> theta <- 0.5 - 0.5*exp(-0.02*20)
> round(trans.mat(theta),3)
           to.IBD=0 to.IBD=1 to.IBD=2
from.IBD=0    0.525    0.399    0.076
from.IBD=1    0.200    0.601    0.200
from.IBD=2    0.076    0.399    0.525
```

It can be shown that the null IBD distribution: $(1/4, 1/2, 1/4)$ is the stationary distribution of such a matrix, i.e., multiplication of the transition matrix on the left by this row vector produces exactly the same vector:

```
> Pr <- c(0.25,0.5,0.25)
> Pr%*%trans.mat(theta)
     to.IBD=0 to.IBD=1 to.IBD=2
[1,]     0.25      0.5     0.25
```

The initial distribution and transition matrices of the Markov process and the conditional distribution of the observations determine the joint distribution of the observations and of the hidden process in a straightforward way. In principle, the determination of the conditional distribution of the hidden process, given the observation is a straightforward application of Bayes' formula. However, a naïve attempt to apply the formulas will face severe computational problems when the number of markers is even moderately large. Indeed, the sample space of the possible paths of the IBD processes over the set of markers grows exponentially fast in powers of three as a function of the number of markers. An attempt to compute directly the posterior distribution of the paths may require the manipulation of an extremely large number of terms and turns out to be impractical if more than a score of markers is considered.

As a remedy, clever algorithms have been developed for computation in HMM scenarios where the unobserved process possesses a Markovian structure and the observations are conditionally independent, given the states of the unobserved process. These algorithms exploit the sequential independence of the components of the process in order to subdivide the task of summing an exponential number of products into a sequence of multiplications of sums

with a fixed number of summands. Below we apply two basic algorithms that were developed for computation in such a setting: The *forward* and the *backward* algorithms.

Denote by $J(t_m)$ the IBD process at the mth marker and by $\Pr(G_m \,|\, j)$ the probability of the observed genotypes at that marker, given that the IBD status is j. Denote the transition probability for the IBD process by $T_{ij} = \Pr(J(t_m) = j \,|\, J(t_{m-1}) = i)$, which also equals $\Pr(J(t_{m-1}) = j \,|\, J(t_m) = i)$, since the process of recombination does not depend on the way we order the markers, but only on the distance between them. In principle T_{ij} will also depend on m when distances between markers vary. The forward algorithm computes recursively the quantity $F_m(j) = \Pr(G_1, G_2, \ldots, G_m, J(t_m) = j)$, which is the joint distribution of the genotypes up to locus t_m and IBD status at that locus. It does so by conditioning on the states of hidden process at the locus t_{m-1} and exploiting the Markovian structure and independence in order to obtain the relation:

$$\begin{aligned} F_m(j) &= \sum_i \Pr(G_1, G_2, \ldots, G_m, J(t_{m-1}) = i, J(t_m) = j) \\ &= \sum_i \left\{ F_{m-1}(i) \times T_{ij} \times \Pr(G_m \,|\, j) \right\} . \end{aligned}$$

Summation in the relation extends over the three possible values of the IBD process at locus t_{m-1}. Applying this relation recursively, starting with the initial relation $F_1(j) = \Pr(G_1 \,|\, j)\Pr(J(t_1) = j)$, allows for the computation of these quantities for all m and j. The number of elements that needs to be manipulated is proportional to the product of the number of markers with the number of possible states of the underlying process (which equals three in our setting).

Denote by $\tilde{m}$ the index of the last marker on the given chromosome. The backward algorithm is used in order to compute the quantity $B_m(j) = \Pr(G_{m+1}, G_{m+2}, \ldots, G_{\tilde{m}} \,|\, J(t_m) = j)$, namely the conditional distribution of the genotypes beyond a locus, given the IBD status at the locus. A recursive relation between the quantities can be identified. This time the relation involves a sum over the states of the process at t_{m+1} and takes the form

$$\begin{aligned} B_m(j) &= \sum_i \Pr(G_{m+1}, \ldots, G_{\tilde{m}}, J(t_{m+1}) = i \,|\, J(t_m) = j) \\ &= \sum_i \left\{ B_{m+1}(i) \times T_{ji} \times \Pr(G_{m+1} \,|\, i) \right\} . \end{aligned}$$

The starting values for the recursion are $B_{\tilde{m}}(j) = 1$.

Let $G = (G_1, \ldots, G_{\tilde{m}})$ be the genetic information over the chromosome for the given pedigree. Since $\Pr(G, J(t_m) = j) = \Pr(G_1, \ldots, G_{\tilde{m}}, J(t_m) = j) = F_m(j)B_m(j)$, it follows from the definition of conditional probabilities that

$$\Pr\big(J(t_m) = j \,|\, G\big) = \frac{F_m(j)B_m(j)}{\sum_i F_m(i)B_m(i)} . \tag{9.10}$$

Consequently, the conditional distribution of IBD, given the genotype information, can be computed at each locus as a function of the F and B quantities.

The functions "`forward`" and "`backward`" apply the forward and backward algorithms in order to compute the forward an backward joint distributions of the genotypes and the IBD states. The first argument to these functions is an array "`G.I`" with the conditional probabilities of the observed genotypes for each of the pedigrees, each of the markers and each of the IBD states. The second and third arguments are the transition matrix of the IBD process and its initial distribution, respectively. The output are arrays that contain the forward and backward probabilities, respectively:

```
> forward <- function(G.I,Tr,Pr)
+ {
+     n.samp <- dim(G.I)[1]
+     n.mark <- dim(G.I)[2]
+     F <- G.I
+     F[,1,] <- sweep(G.I[,1,],2,Pr,"*")
+     for (i in 2:n.mark)
+     {
+         F[,i,] <- G.I[,i,]*(F[,i-1,]%*%Tr)
+         S <- F[,i,1] + F[,i,2] + F[,i,3]
+         F[,i,] <- sweep(F[,i,],1,S,"/")
+     }
+     return(F)
+ }
> backward <- function(G.I,Tr,Pr)
+ {
+     n.samp <- dim(G.I)[1]
+     n.mark <- dim(G.I)[2]
+     B <- G.I
+     B[,n.mark,] <- 1
+     for (i in seq(n.mark-1,1))
+     {
+         B[,i,] <- (G.I[,i+1,]*B[,i+1,])%*%t(Tr)
+         S <- B[,i,1] + B[,i,2] + B[,i,3]
+         B[,i,] <- sweep(B[,i,],1,S,"/")
+     }
+     return(B)
+ }
```

Note that in each iteration of the evaluation, the currently computed quantities in `F` and `B` are re-scaled to sum to one. This re-scaling increases the numerical stability of the algorithm, which would otherwise involve the manipulation of terms that become vanishingly small as the algorithm progresses. Round off errors would have been a serious concern if that were the case. Owing to the re-scaling, the terms are no longer the probabilities per se, but

are only proportional to such probabilities. Nonetheless, the constants of proportionality do not depend on the IBD status at a locus and are therefore canceled out of both the numerator and the denominator when (9.10) is applied in order to obtain the target distribution. The actual computation of the conditional distribution of the states, given the genotypes, is carried out in the function "`marginal.post`". This function takes as input the output arrays of the functions "`forward`" and "`backward`" and produces an array of the same type with the posterior probabilities of the states:

```
> marginal.post <- function(F,B)
+ {
+     P <- F*B
+     S <- P[,,1]+P[,,2]+P[,,3]
+     P <- sweep(P,1:2,S,"/")
+     return(P)
+ }
```

9.6.3 Statistical Properties of Genome Scans

With the tools developed for the simulation of random pedigrees and a function for the computation of the estimated identity-by-descent probabilities from the genotypic information, we can start investigating the statistical properties of mapping in the more realistic setting of partial information. Let us initiate our investigation by the determination of the expectation and covariance properties of the scanning statistic under both the null and the alternative hypotheses. Later, we will consider Gaussian processes with the same covariance and mean structure and obtain significance thresholds and power curves. Our investigation will include several inter-marker spacings.

The first simulation is conducted under the null distribution. We simulate 10^5 independent pedigrees and use them in order to compute the covariance and mean structure of the reconstructed IBD process. We also assess the accuracy of the reconstruction at a central locus. The simulation is split into 100 batches in order to avoid running out of memory:

```
> n.rep <- 10^2
> n.ped <- 10^3
> Delta <- c(35,20,10,5,1)
> ibd.est.null <- matrix(nrow=3,ncol=length(Delta))
> colnames(ibd.est.null) <- paste("Delta=",Delta,sep="")
> rownames(ibd.est.null) <- c("mean","var","mse")
> cor.ibd <- vector(mode="list",length=length(Delta))
> names(cor.ibd) <- paste("Delta=",Delta,sep="")
> cor.est <- cor.ibd
> P <- geno.given.ibd()
> for(i in 1:length(Delta))
```

```
+ {
+     markers <- seq(0,140,by=Delta[i])
+     n.mark <- length(markers)
+     locus <- ceiling(n.mark/2)
+     theta <- 0.5 - 0.5*exp(-0.02*Delta[i])
+     Tr <- trans.mat(theta)
+     fa <- list(pat=matrix(1,n.ped,n.mark),
+                mat=matrix(2,n.ped,n.mark))
+     mo <- list(pat=matrix(3,n.ped,n.mark),
+                mat=matrix(4,n.ped,n.mark))
+     ibd.est <- ibd <- NULL
+     G.I <- array(dim=c(n.ped,n.mark,3))
+     for (rep in 1:n.rep)
+     {
+         sib1 <- mating(fa,mo,markers)
+         sib2 <- mating(fa,mo,markers)
+         geno <- ped.geno(sib1,sib2)
+         for(k in 1:3) G.I[,,k] <-
+               matrix(P[geno,k],n.ped,n.mark)
+         F.P <- forward(G.I,Tr,Pr)
+         B.P <- backward(G.I,Tr,Pr)
+         I.G <- marginal.post(F.P,B.P)
+         ibd.est <- rbind(ibd.est,2*I.G[,,3]+I.G[,,2])
+         ibd <- rbind(ibd,(sib1$pat == sib2$pat)+
+                 (sib1$mat == sib2$mat))
+     }
+     ibd.est.null["mean",i] <- mean(ibd.est[,locus])
+     ibd.est.null["var",i] <- var(ibd.est[,locus])
+     ibd.est.null["mse",i] <-
+                mean((ibd[,locus]-ibd.est[,locus])^2)
+     cor.ibd[[i]] <- cor(ibd)
+     cor.est[[i]] <- cor(ibd.est)
+ }
```

Observe that mean, variance, and mean-square distance between the reconstructed and actual IBD processes are stored in a matrix called "`ibd.est.null`". The columns of this matrix correspond to the different inter-marker spacings. The correlation matrices are stored in a list named "`cor.est`". For later comparison, we also store the correlation structure of the actual IBD process in the list "`cor.ibd`".

Let us examine the mean, the variance, and the quality of reconstruction under the null distribution:

```
> round(ibd.est.null,3)
     Delta=35 Delta=20 Delta=10 Delta=5 Delta=1
mean    1.001    1.002    1.001   1.002   0.998
var     0.126    0.155    0.217   0.292   0.425
mse     0.374    0.343    0.284   0.208   0.075
```

Several insights emerge. First, it can be seen that the expected value of the estimated IBD is about equal to the expectation of the true IBD, which is one. As a matter of fact, it can be shown mathematically that the expectations of the estimated and the true IBD coincide. This follows from the fact that the estimated expression is a conditional expectation and the mathematical fact that the expectation of a random variable is equal to the expectation of its conditional expectation. (Symbolically, $\mathrm{E}(J) = \mathrm{E}[\mathrm{E}(J|G)]$.)

Second, one can conclude that the variance of the estimated IBD is less than the variance of the actual IBD, which is equal to 1/2. Moreover, the denser the markers are, the closer the variance is to 1/2. Mathematically the relation between variances of the actual IBD J and the estimated IBD, which we denote by $\hat{J}$, is given by the relation

$$\mathrm{var}(J) = \mathrm{var}(\mathrm{E}(J \,|\, G)) + \mathrm{E}(\mathrm{var}(J \,|\, G)) = \mathrm{var}(\hat{J}) + \mathrm{E}(\mathrm{var}(J \,|\, G)) > \mathrm{var}(\hat{J}).$$

The more informative G is, the more similar $\hat{J}$ is to J and the closer their variances are to each other.

Third, a closer examination of the numbers in the second and third rows reveals that their sum equals one-half – the variance of the actual IBD statistic. In mathematical terms one can express this relation in the form:

$$\mathrm{E}[(\hat{J} - J)^2] = \mathrm{E}[\mathrm{var}(J \,|\, G)].$$

Substituted into the previous equation, this relation shows that the closer the variance of the reconstructed IBD is to the variance of the actual IBD, the more accurate the reconstruction is, as measured by its mean squared error.

In the computation of a scanning statistic, standardization is carried out with the standard deviation of the estimated IBD. The resulting statistic has a zero mean and a unit variance under the null distribution. The Gaussian limiting distribution of the process of statistics is determined by the correlation structure. In Fig. 9.3 the correlation function between a statistic computed for a central marker and the statistics computed at the flanking markers is plotted. The spacing between markers is 10 cM. The code that produced the figure is:

```
> markers <- seq(0,140,by=10)
> plot(markers,cor.ibd$"Delta=10"[8,],pch=19,ylab="cor.")
> points(markers,cor.est$"Delta=10"[8,])
> legend(1,0.99,legend=c("Actual","Estimated"),pch=c(19,1))
```

The correlation values for the actual IBD process are plotted in *solid black* and the values for the estimated IBD process are plotted in *black and white*.

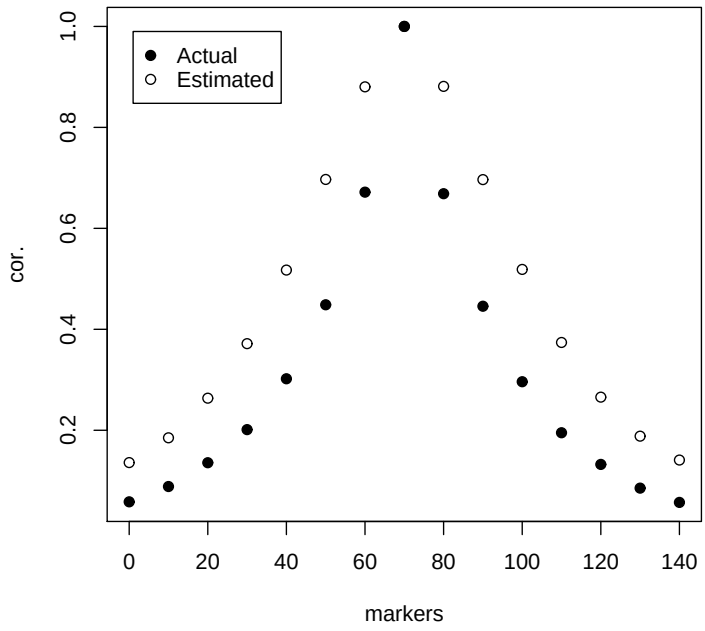

Fig. 9.3. Correlation functions between markers for IBD processes (for an inter-markers spacing of $\Delta = 10$ cM).

Observe that in general the correlations for the estimated process are larger than for the actual IBD process. Indeed, the same set of genotypes is used in order to infer the IBD status at the different markers. This increases the correlation beyond the correlation that results from the recombination process. A byproduct of this increase in the correlation will be a decrease in the size of the threshold that should be used to ensure a given significance level in a genome scan. Indeed, the first step in the investigation of the statistical properties of a genome scan involves a slight readjustment of the threshold to the new setting. We will get to that below.

Before considering the process of test statistics let us examine the properties of the reconstructed IBD when a susceptibility gene is present. We consider the same inter-markers spacings and assume that the locus is perfectly linked to the central marker. We assume an additive model for the trait with $p = 0.1$, $g_0 = 0.05$, and $\alpha = 0.225$. The means, variances, and mean square errors are computed for the marker linked to the trait locus. The simulation applies, with some obvious modifications, the same code as before:

```
> ibd.prob <- DistIBD(0.1,0.05,0.275,0.5)
```

```
> ibd.est.alt <- ibd.est.null
> for(i in 1:length(Delta))
+ {
+     markers <- seq(0,140,by=Delta[i])
+     n.mark <- length(markers)
+     qtl <- ceiling(n.mark/2)
+     theta <- 0.5 - 0.5*exp(-0.02*Delta[i])
+     Tr <- trans.mat(theta)
+     fa <- list(pat=matrix(1,n.ped,n.mark),
+                mat=matrix(2,n.ped,n.mark))
+     mo <- list(pat=matrix(3,n.ped,n.mark),
+                mat=matrix(4,n.ped,n.mark))
+     G.I <- array(dim=c(n.ped,n.mark,3))
+     ibd.est <- ibd <- NULL
+     for (rep in 1:n.rep)
+     {
+         inhe.qtl <- inhe.vector(ibd.prob,n.ped)
+         sib1 <- mating(fa, mo, markers,qtl=markers[qtl],
+                 inhe.fa=inhe.qtl[,"sib1.pat"],
+                 inhe.mo=inhe.qtl[,"sib1.mat"])
+         sib2 <- mating(fa, mo, markers,qtl=markers[qtl],
+                 inhe.fa=inhe.qtl[,"sib2.pat"],
+                 inhe.mo=inhe.qtl[,"sib2.mat"])
+         geno <- ped.geno(sib1,sib2)
+         for(k in 1:3) G.I[,,k] <-
+               matrix(P[geno,k],n.ped,n.mark)
+         F.P <- forward(G.I,Tr,Pr)
+         B.P <- backward(G.I,Tr,Pr)
+         I.G <- marginal.post(F.P,B.P)
+         ibd.est <- c(ibd.est,2*I.G[,qtl,3]+I.G[,qtl,2])
+         ibd <- c(ibd,((sib1$pat == sib2$pat)+
+             (sib1$mat == sib2$mat))[,qtl])
+     }
+     ibd.est.alt["mean",i] <- mean(ibd.est)
+     ibd.est.alt["var",i] <- var(ibd.est)
+     ibd.est.alt["mse",i] <- mean((ibd-ibd.est)^2)
+ }
```

The matrix "`ibd.est.alt`" stores the mean, the variance, and the mean square error for the marker fully linked with the trait:

```
> round(ibd.est.alt,3)
     Delta=35 Delta=20 Delta=10 Delta=5 Delta=1
mean    1.042    1.052    1.073   1.096   1.144
var     0.116    0.147    0.218   0.297   0.420
mse     0.353    0.325    0.267   0.193   0.068
```

As expected, the mean of the estimated IBD is elevated. The elevation is more apparent when markers are denser. The variance and mean square error are, however, hardly affected by the change in distribution from unlinked to linked. This last observation is consistent with the mathematics of local alternatives, where one varies the mean but not the covariance structure.

It is more convenient to interpret the effect that the missing information may have on the statistical power by considering the noncentrality parameter. This parameter is equal to the difference between the alternative and the null expectations, multiplied by the square root of the sample size and divided by the null standard deviations. If we take, for example, a trial with 400 pedigrees, then we obtain

```
> n <- 400
> ncp.ibd <- sqrt(2*n)*(sum(ibd.prob*0:2)-1)
> ncp.app <- ncp.ibd*sqrt(ibd.est.null["var",]/0.5)
> ncp.sim <- sqrt(n)*(ibd.est.alt["mean",]-1)/
+     sqrt(ibd.est.null["var",])
> ncp <- rbind(ncp.sim,ncp.app)
> round(ncp.ibd,3)
[1] 4.744
> round(ncp,3)
        Delta=35 Delta=20 Delta=10 Delta=5 Delta=1
ncp.sim    2.351    2.640    3.139   3.552   4.412
ncp.app    2.383    2.638    3.128   3.626   4.372
```

The noncentrality parameter for a fully informative marker with complete linkage is 4.744. The same parameter for the same marker when only partial information is available was computed in two different ways. In the variable "`ncp.sim`" it was computed directly based on the results of the simulations. In the variable "`ncp.app`" it was approximated using (9.11) given below.

Looking at the numbers we see that the noncentrality parameter is severely deflated if the inter-marker spacing is more than 5 cM. If the spacing is 1 cM or less, then one recovers most of the noncentrality parameter. Another observation is the similarity between the actual noncentrality values as computed by simulations and the following approximation for these values. This approximation takes the general form:

$$\mathrm{E}(\hat{Z}_t) \approx \mathrm{cor}\big(\hat{J}(t), \hat{J}(\tau)\big) \times \left[\frac{\mathrm{var}(\hat{J}(\tau))}{\mathrm{var}(J(\tau))}\right]^{1/2} \times \mathrm{E}(Z_\tau)\,, \tag{9.11}$$

where $\mathrm{E}(Z_\tau) = \xi$ is the noncentrality parameter which was considered in the previous section and computed at the trait locus under the assumption of completely informative markers, while the statistic $\hat{Z}_t$ is the test statistic computed at a marker and based on the reconstructed IBD. The term $\mathrm{var}(J(\tau))$ is the variance of the actual IBD process, 1/2 in this case, and $\mathrm{var}(\hat{J}(t))$ is the variance of the reconstructed IBD, computed at the marker.

The term $\mathrm{cor}\big(\hat{J}(t), \hat{J}(\tau)\big)$ corresponds to the correlations between elements of the reconstructed process.

The noncentrality parameter conveniently decomposes into three factors. The rightmost factor measures the contribution of the genetic effect, combined with the sample size. The central factor measures the reduction in the noncentrality parameter due to missing information. The leftmost factor measures the effect of using a marker that is only partially correlated with the trait locus.

Although we learn something from examining the noncentrality parameter, we now carry out a more comprehensive investigation of the statistical properties of genome scans when only the siblings are available for genotyping. The first stage involves finding the appropriate thresholds for the different inter-marker spacings. For the actual IBD process we use the analytical approximations that were developed in Chap. 4. For the reconstructed IBD we use simulations. Observe that the correlation structure of a scanning process formed by the summation of independent copies of reconstructed IBD processes is the same at the correlation structure of a single such process. Hence, we can use the correlation matrices that we found for the reconstructed processes as inputs for the function that simulates the Gaussian processes.

```
> library(MASS)
> Delta <- c(35,20,10,5,1)
> z <- matrix(nrow=2,ncol=length(Delta))
> colnames(z) <- paste("Delta=",Delta,sep="")
> rownames(z) <- c("ibd","ibd.est")
> n.rep <- 10^2
> n.iter <- 10^4
> for(d in 1:length(Delta))
+ {
+     z["ibd",d] <- uniroot(OU.approx,c(3,4),beta=0.04,
+         Delta=Delta[d],length=140*23,chr=23,center=0.05,
+         test="one-sided")$root
+     n.mark <- 140/Delta[d] + 1
+     Z.max <- NULL
+     for(i in 1:n.rep)
+     {
+         Z <- mvrnorm(n.iter,rep(0,n.mark),cor.est[[d]])
+         Z.max <- c(Z.max,apply(Z,1,max))
+     }
+     z["ibd.est",d] <- sort(Z.max)[n.rep*n.iter*(1-0.05/23)]
+ }
> round(z,3)
        Delta=35 Delta=20 Delta=10 Delta=5 Delta=1
ibd        3.338    3.461    3.603   3.722   3.907
ibd.est    3.322    3.425    3.515   3.587   3.742
```

As expected from our comparison of correlations, the threshold levels for the reconstructed IBD are slightly smaller than the threshold levels for the actual IBD process. We will use these lower thresholds in order to determine the power.

For the actual IBD process we use the analytical approximation that was developed in Chapter 6. For the estimated IBD we again use simulations. Motivated by the theory of local alternatives, we simulate the process under the alternative distribution using the same covariance structure that was used under the null. The mean function is computed based on the approximation (9.11) and is added to each row of the random part using the function "`sweep`":

```
> xi <- seq(3,6,by=0.1)
> n.rep <- 10
> power.ibd <- matrix(nrow=length(xi),ncol=length(Delta))
> colnames(power.ibd) <- names(z)
> power.est <- power.ibd
> for (d in 1:length(Delta))
+ {
+     power.ibd [,d] <-
+         power.marker(z["ibd",d],0.04,Delta[d],xi)
+     n.mark <- 140/Delta[d] + 1
+     qtl <- ceiling(n.mark/2)
+     rho <- sqrt(ibd.est.null["var",d]/0.5)*
+         ((cor.est[[d]])[qtl,])
+     Z <- mvrnorm(n.iter,rep(0,n.mark),cor.est[[d]])
+     for (i in 1:length(xi))
+     {
+         ncp <- xi[i]*rho
+         Z1 <- sweep(Z,2,ncp,"+")
+         power.est[i,d] <-
+             mean(apply(Z1,1,max) >= z["ibd.est",d])
+     }
+ }
```

Let us plot the power functions:

```
> plot(range(xi),c(0,1),type="n",xlab="xi",ylab="power")
> gr <- gray(0.75*(1:length(Delta))/length(Delta))
> for(d in 1:length(Delta))
+     {
+         lines(xi,power.ibd[,d],col=gr[d])
+         lines(xi,power.est[,d],col=gr[d],lty=2)
+     }
> legend(xi[1],1,bty="n",legend=colnames(z),
+     lty=rep(1,ncol(z)),col=gr)
```

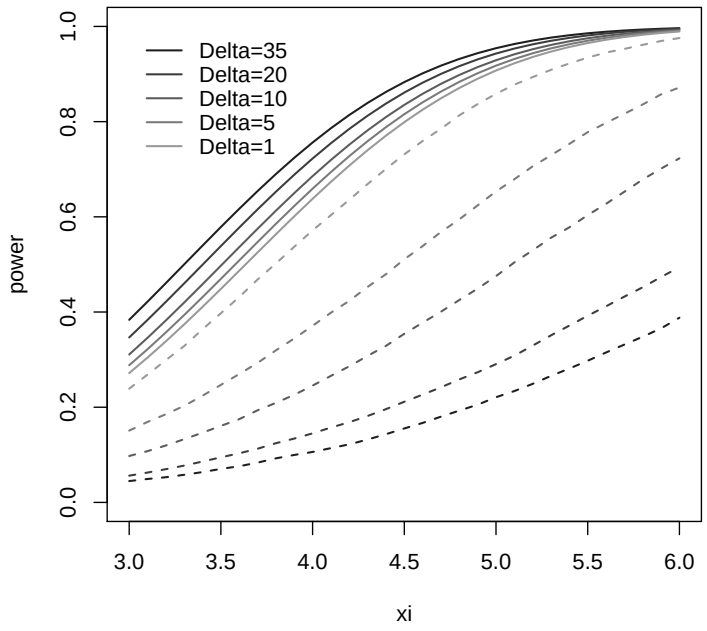

Fig. 9.4. Power functions for an IBD process (*solid line*) and an inferred process (*broken line*). The QTL in complete linkage with the central marker. Bi-allelic markers are used.

The output is presented in Fig. 9.4. Observe the dramatic effect on the power of failing to reconstruct the IBD process accurately enough. When markers are placed one cM apart, then most of the power is retained. However, a substantial fraction of the power is lost when markers are set 5 cM apart and becomes worse as the markers become more spread apart.

The exact same programs that were used to generate Fig. 9.4, which refers to SNPs, can be used in order to generate power curves for SSR markers. Consider, for example, markers with five uniformly distributed alleles. The only changes that are needed are the replacement of line "`P <- geno.given.ibd()`" by the line "`P <- geno.given.ibd(rep(1/5,5))`" in one location and the line "`geno <- ped.geno(sib1,sib2)`" by the line "`geno <- ped.geno(sib1,sib2,rep(1/5,5))`" in two locations.

After making these changes and rerunning the programs one gets new significance thresholds and new power curves that are given in Fig. 9.5:

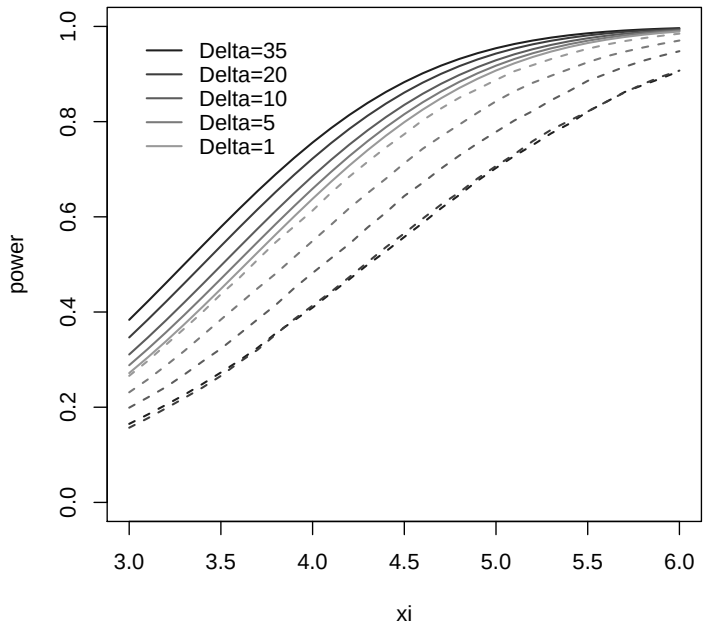

Fig. 9.5. Power functions for an IBD process (*solid line*) and an inferred process (*broken line*). The QTL in full linkage with the central marker. Uniformly distributed five-allelic markers are used.

	Delta=35	Delta=20	Delta=10	Delta=5	Delta=1
ibd	3.338	3.461	3.603	3.722	3.907
ibd.est	3.321	3.442	3.563	3.663	3.841

Clearly, for markers separated by more than a few cM, the SSRs provide more information than the SNPs. Indeed, SSRs at 5 cM appear to be roughly comparable to SNPs at 1 cM. If markers are multi-allelic or if parents are also genotyped, then one can recover more information and mitigate to some extent the effect of separated markers. Nevertheless, when one takes account of the effect of a value of β twice as large as one finds in backcross or intercross designs and the additional loss of information due to incompletely informative markers, one should plan to place markers substantially closer in human than in experimental genetics. With microsatellite markers it is common to use an inter-marker spacing of about 10 cM, and 5 cM would be substantially better. For SNPs, which are generally less informative but more common than microsatellites, the marker spacing should be about 1 cM, or less.

9.7 Bibliographical Comments

The idea of using affected sib pairs to map disease genes goes back to Penrose [58]. James [42] observed that the regression models used for quantitative traits could also be used for qualitative traits if one treated the (unobserved) penetrance of a qualitative trait as a quantitative phenotype. Risch [65] provided an elegant theoretical framework for single marker applications. For discussion of genome scans, see Feingold, Brown, and Siegmund [28] and Lander and Kruglyak [48].

The standard paradigm of parametric linkage analysis is due to Morton [56], and is described in detail by Ott [57].

Considerable effort has gone into the problem of reconstructing IBD relations from marker data. One of the first and still widely used programs for this purpose is based on the Elston-Stewart [24] algorithm, implemented in the software SAGE. As befits its vintage, this algorithm is particularly adept at dealing with large pedigrees and a relatively small number of markers. Two widely used and freely available programs that are built around a hidden Markov model of the kind discussed in the text are GENEHUNTER (Kruglyak *et al.* [46]) and MERLIN (Abecassis *et al.* [2]). Both of these programs can handle essentially any number of markers, but have problems in dealing with large pedigrees. ALLEGRO [33] began its existence as something of an offspring of GENEHUNTER, but lately [34] seems to have developed an independent life of its own.

These programs also provide suites of statistical methods that utilize the IBD reconstruction for gene mapping. The reconstruction is computationally intensive, so the programs should be optimized; and since the goals are reasonably clear, one would usually want to use programs prepared by professionals. The statistical applications, on the other hand, are comparatively easy to program using available statistical software, and the appropriate methods are not so generally agreed on. Hence one might want to use the IBD reconstruction of, for example, GENEHUNTER or MERLIN as a preliminary step to one's own statistical analysis.

Problems

9.1. (a) Plot the mean function of the Z statistic for an additive model with $g_0 = 0.05$, for various values of p, α, and n.
(b) For inter-marker spacing of $\Delta = 5$ cM, plot the power function for various values of p, α, and n.
(c) For $p = 0.1$ and $\alpha = 0.4$, how large a value of n is required to have 90% power to detect a gene perfectly linked to a marker?

9.2. Use (9.4) to find the probabilities of IBD for an additive model with $g_0 = 0$. Note, in particular, that these probabilities depend only on p and

not on the penetrances $g_1 = \alpha,\ g_2 = 2\alpha$. Do the same for the dominant model $g_0 = 0,\ g_1 = g_2 = \alpha$. Make numerical comparisons of the noncentrality parameters of $(N_2 - N_0)/(n/2)^{1/2}$ for the two models for different values of p. Do you find your numerical results surprising?

9.3. The probabilities of IBD for the recessive model $0 = g_0 = g_1 < g_2$ are given by

$$\pi_0 = \frac{p^2}{(1+p)^2}, \quad \pi_1 = \frac{2p}{(1+p)^2}, \quad \pi_2 = \frac{1}{(1+p)^2},$$

where p is the frequency of the allele D in the population.

(a) Write an `R` function that computes the power to detect a susceptibility gene as a function of the relevant parameters.
(b) Plot the power function for various values these parameters.

9.4. The statistic $Z_1 = (N_2 - N_0)/(n/2)^{1/2}$ is designed for an additive model. For a recessive model, where $\sigma^2_{\mathrm{D}} \gg \sigma^2_{\mathrm{A}}$, an alternative is $Z_2 = (N_2 - n/4)/(3n/16)^{1/2}$.

(a) Explain why this is a reasonable statistic (for example, by calculating its noncentrality parameter).
(b) For the additive and recessive models of the preceding two problems, for which $g_0 = 0$, compare the noncentrality parameters of Z_1 and Z_2 as functions of p. For which values of p does Z_2 have a substantially larger noncentrality parameter than Z_1? (See Chap. 11 for additional discussion of alternative statistics when the dominance variance may be important.)

9.5. Consider the model (2.3) for the penetrance y, where e is assumed to contain environmental effects and possibly the effect of other genes that are unlinked to the trait locus at τ. Since two siblings are expected to share genetic material and are also likely to share a common environment, we allow for a covariance $r = \mathrm{cov}(e_1, e_2)$. How should the assumption (9.1) be changed? What is the effect of this generalization on (9.2) and (9.4)?

9.6. Simulate significance thresholds at various inter-marker distances when markers are only partially informative and compare these thresholds to those obtained (either by simulation or by theoretical approximations) for fully informative markers. Now make similar comparisons of the power of genome scans based on the appropriate thresholds.

9.7. An alternative to the additive model, which is also particularly tractable, is the multiplicative relative risk model. The penetrances are given by $g_0 > 0$, $g_1 = g_0 R$, $g_2 = g_0 R^2$, where $R > 1$ is called the relative risk. Assuming Hardy-Weinberg equilibrium, for a sib pair find expressions for Pr(Both affected) and the conditional probability Pr(Both affected$|J(\tau) = j$) for $j = 0, 1, 2$. Generalize this model to consider two (unlinked) genes acting multiplicatively between loci.

Remark 9.2. It is possible to reduce this to the canonical form given in Sect. 9.1 and apply the results given in the text; but the properties of this model are sufficiently simple that one can also carry out the desired calculations from first principles.

9.8. Let $\sigma_{\rm D}^2 = 0$ in (9.4). The log likelihood function for the data (N_0, N_1, N_2) parametrized by $\check{\alpha}/Q_2$ is $\ell(\check{\alpha}/Q_2) = N_2 \log(1 + \check{\alpha}/Q_2) + N_0 \log(1 - \check{\alpha}/Q_2)$. Show that the statistic Z introduced in the text is the score statistic for testing the hypothesis $H_0 : \check{\alpha} = 0$, which is of the form $\dot{\ell}(0)/\{\mathrm{E}_0[\dot{\ell}(0)]^2\}^{1/2}$. (We use the dot notation for derivatives. See Chap. 1 for a discussion of the score statistic.)

9.9. Let $\pi_i(\theta) = \Pr(J(t) = i|A)$ for a marker t at a recombination fraction θ from a trait locus τ. Show that $\pi_i(\theta)$ is of the same form as (9.4), but with $\check{\alpha}$ replaced by $\check{\alpha}(1 - 2\varphi)$ and $\check{\delta}$ replaced by $\check{\delta}(1 - 2\varphi)^2$.

9.10. The affected sib pair method is based on the observation that affected sibs are likely to share an excess of alleles IBD at a locus that increases susceptibility to the trait. A sib pair of whom one affected and one unaffected is likely to share a deficit of alleles IBD at a trait locus. Assuming there is no dominance variance, develop a model for using affected/unaffected sib pairs for gene mapping. Find the score statistic or an otherwise reasonable statistic and evaluate its noncentrality parameter. Can you give conditions where these sib pairs would be as useful as affected sib pairs? Does it seem plausible that these conditions might sometimes be satisfied?

9.11. Suppose that our sample consists of n_1 affected/unaffected sib pairs, as in the preceding problem, and n_2 affected sib pairs. Assuming a parametric model, where we regard the penetrances and allele frequencies as known, and the validity of (9.8), propose an approximate Gaussian log likelihood ratio statistic to combine these two kinds of sib pairs. Observe that this involves a specific linear combination of the two Z statistics, with weights that depend on the hypothesized values of the noncentrality parameters. (This issue is discussed again in Chap. 11.)

9.12. Give an analysis along the lines of Sects. 9.1–9.3 for samples of (i) an affected grandparent and grandchild (ii) affected half-sibling pairs, (iii) affected first cousins pairs.

Remark 9.3. Determining the recombination parameter for first cousins involves a more difficult argument than the other two cases.

9.13. Compute the probability distribution of the total number of alleles shared IBD by a trio of siblings.

9.14. Show that (9.3) can be re-written in the form

$$\mathrm{E}(y_1 y_2|J) = \mathrm{E}(y_1 y_2) + (J - 1)\check{\alpha} - (I_{\{J=1\}} - 1/2)\check{\delta} \, .$$

9.15. Recall that a trait is called recessive if $\delta = -\alpha$. Consider a child whose parents are related to each other, e.g., siblings or first cousins. Let F denote the coefficient of relatedness of the parents. Using the model (2.3) together with the assumption that $\mathrm{E}(e) = 0$, show that the probability the child is affected is $Q_1 = m + F\sigma_{\mathrm{D}}$. In particular, the probability that an inbred child is affected is larger than the probability m that a child of unrelated parents is affected. An individual is said to be homozygous by descent (HBD) at the locus t if the two alleles at that locus are IBD. Hence the the probability of HBD at a random locus is the coefficient of relatedness of the parents. Explain how you could use a sample of inbred affected individuals, e.g., a sample of affected children of first cousins, to map a recessively acting gene. What would be the noncentrality parameter at a trait locus in a large sample in terms of the sample size n, Q_1, F, and σ_{D}^2? How would your analysis change if you want to consider the possibility of a second gene, which is unlinked to and does not interact with the first gene?