Nonparametric linkage (NPL) analysis

Mapping of disease genes for Mendelian diseases:

- Sample: one or a few large pedigrees
- Analysis: parametric linkage analysis
- Success story: disease loci for more than 1,500 Mendelian diseases have been identified

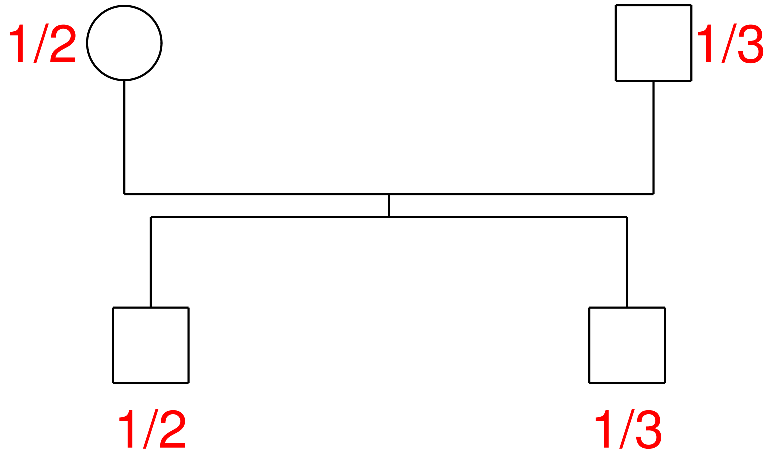
Mapping of disease genes for genetically complex diseases:

- Sample: large collection of small pedigrees (often: affected sib pairs)
- Analysis: nonparametric linkage analysis (allele sharing methods)
- Success story: much less impressive until now

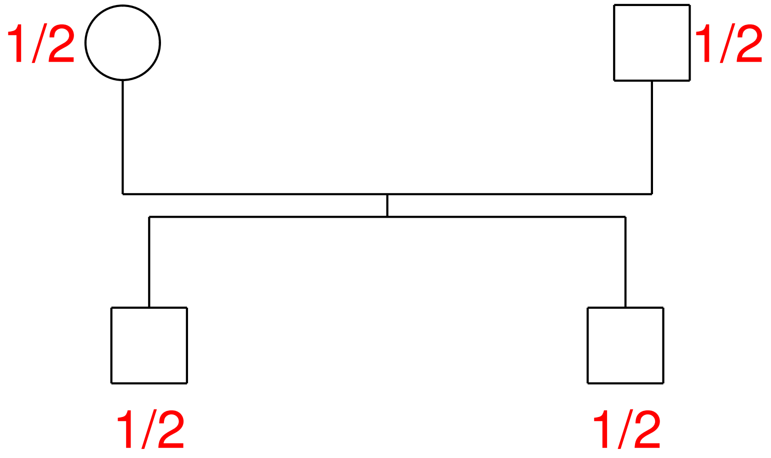
Allele sharing: IBS and IBD

IBS: Two alleles are said to be *identical by state* (IBS) if they are of the same kind

IBD: Two alleles are said to be *identical by descent* (IBD) if both of them are copies of the same ancestral allele



IBS=1
IBD=0



IBS=2
IBD=0 or 2

Distribution of the IBD score in sib pairs

Assumption:
Each parent transmits each of his or her allele with probability 1/2

Genotype of					
father	mother	child 1	child 2	probability	IBD score
1/2	3/4	1/3	1/3	1/16	2
		1/3	1/4	1/16	1
		1/3	2/3	1/16	1
		1/3	2/4	1/16	0
		1/4	1/3	1/16	1
		1/4	1/4	1/16	2
		1/4	2/3	1/16	0
		1/4	2/4	1/16	1
		2/3	1/3	1/16	1
		2/3	1/4	1/16	0
		2/3	2/3	1/16	2
		2/3	2/4	1/16	1
		2/4	1/3	1/16	0
		2/4	1/4	1/16	1
		2/4	2/3	1/16	1
		2/4	2/4	1/16	2

$$\Rightarrow P(\mathrm{IBD}=2)=P(\mathrm{IBD}=0)=1/4,\ P(\mathrm{IBD}=1)=1/2$$

Distribution of the IBD score in sib pairs

Exercise:

1. Provide a more elegant argument for showing that

   $P(\mathrm{IBD} = 2) = P(\mathrm{IBD} = 0) = 1/4$, $P(\mathrm{IBD} = 1) = 1/2$

   (Hint: Binomial distribution)

2. Show that the expectation of the IBD score in sib pairs is 1 and the variance of the IBD score is 1/2.

Distribution of the IBD score in affected sib pairs

For $k \in \{0, 1, 2\}$, let $z_k = P(IBD = k)$ and $z = (z_2, z_1, z_0)$.

- Marker and disease locus unlinked (i.e., recombination fraction between marker and disease locus $= 1/2$):

  Distribution of IBD scores in affected sib pairs is identical to the distribution of IBD scores in sib pairs, i.e., $z = (1/4, 1/2, 1/4)$.

- Marker and disease locus linked (i.e., recombination fraction between marker and disease locus $< 1/2$):

  Distribution of IBD scores in affected sib pairs is different from $(1/4, 1/2, 1/4)$.

Distribution of the IBD score in affected sib pairs

(z_2, z_1, z_0) at the disease locus depends on the disease model:

- single locus disease model:

$$\begin{aligned} z_2 &= \frac{1}{4} + \frac{V_A/2 + 3V_D/4}{4(K_P^2 + V_A/2 + V_D/4)} \\ z_1 &= \frac{1}{2} - \frac{V_D/2}{4(K_P^2 + V_A/2 + V_D/4)} \\ z_0 &= \frac{1}{4} - \frac{V_A/2 + V_D/4}{4(K_P^2 + V_A/2 + V_D/4)} \end{aligned}$$

with

$$\begin{aligned} V_A &= 2p(1-p)\,[p(f_2 - f_1) + (1-p)(f_1 - f_0)]^2 \qquad \text{(additive variance)} \\ V_D &= p^2(1-p)^2\,[f_2 - 2f_1 + f_0]^2 \qquad \text{(dominant variance)} \\ K_P &= p^2 f_2 + 2p(1-p)f_1 + (1-p)^2 f_0 \qquad \text{(disease prevalence)} \end{aligned}$$

- more complex disease model:

  (z_2, z_1, z_0) at the disease locus can be calculated numerically

Distribution of the IBD score in affected sib pairs

Assume that a marker locus is linked to a disease locus at recombination fraction θ. Let $z^D = (z_2^D, z_1^D, z_0^D)$ denote the distribution of the IBD scores in affected sib pairs at the disease locus.

Goal: Calculation of the distribution $z^M = (z_2^M, z_1^M, z_0^M)$ of the IBD scores in affected sib pairs at the marker locus.

Distribution of the IBD score in affected sib pairs

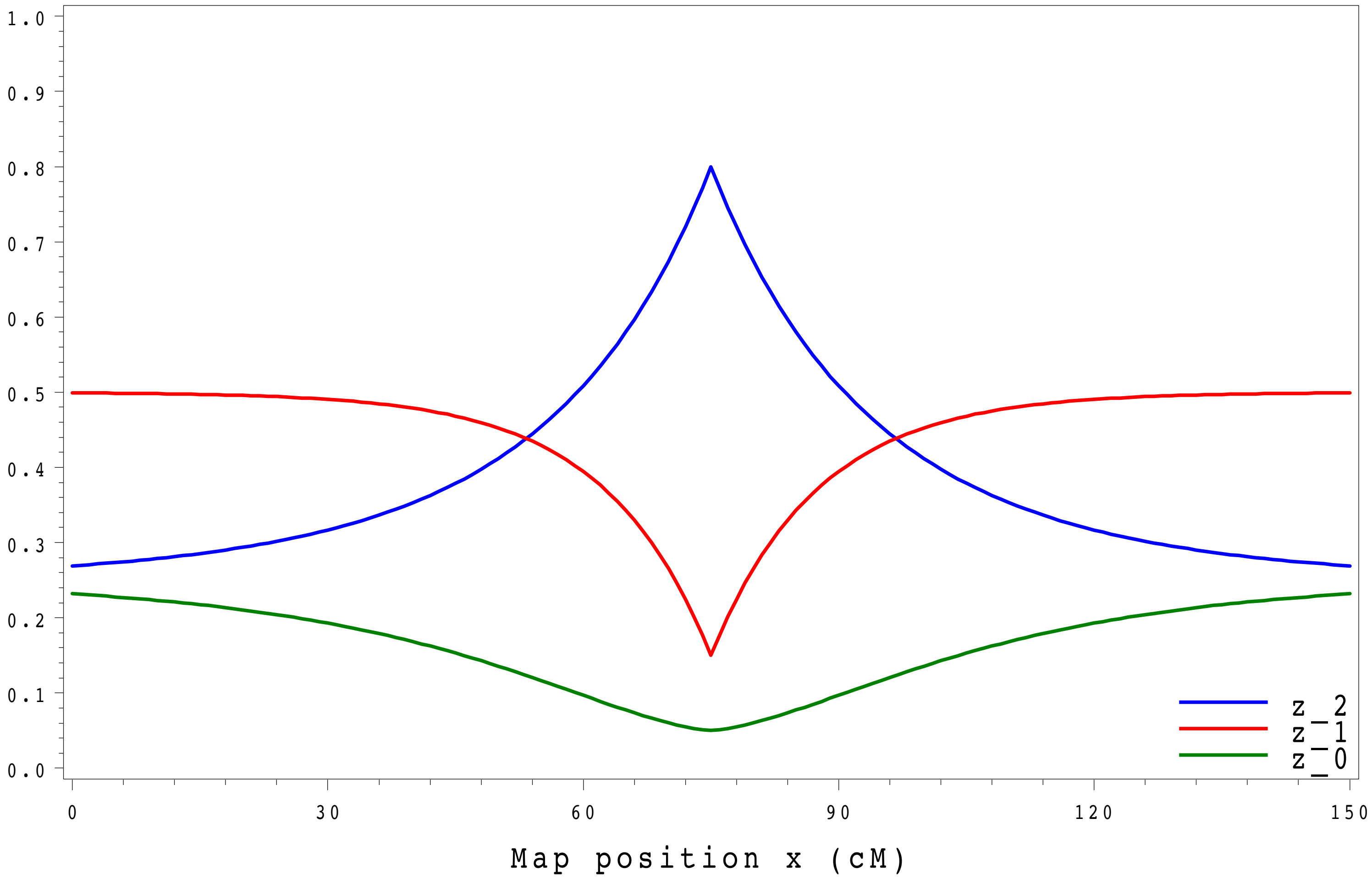
Solution: Let $W_f^D = 1$ (or =0), if the two alleles at the disease locus transmitted by the father are IBD (or not IBD). Let W_f^M be defined analogously for the alleles at the marker locus. Finally, W_m^D and W_m^M are the corresponding random variables for the mother. Then,

$$P(W_f^D = W_f^M) = P(W_m^D = W_m^M) = \theta^2 + (1-\theta)^2 =: \phi,$$

$$P(W_f^D \neq W_f^M) = P(W_m^D \neq W_m^M) = 2\theta(1-\theta) = 1 - \phi$$

$$\begin{aligned} \Rightarrow \quad z_2^M &= \phi^2 z_2^D + \phi(1-\phi) z_1^D + (1-\phi)^2 z_0^D \\ z_1^M &= 2\phi(1-\phi) z_2^D + \left[\phi^2 + (1-\phi)^2\right] z_1^D + 2\phi(1-\phi) z_0^D \\ z_0^M &= (1-\phi)^2 z_2^D + \phi(1-\phi) z_1^D + \phi^2 z_0^D \end{aligned}$$

Distribution of the IBD score in affected sib pairs



z_2 and z_1 for different loci along a chromosome. At the disease locus, positioned at 75 cM, $z_2 = 0.8$ and $z_1 = 0.15$.

ASP tests for a completely informative marker

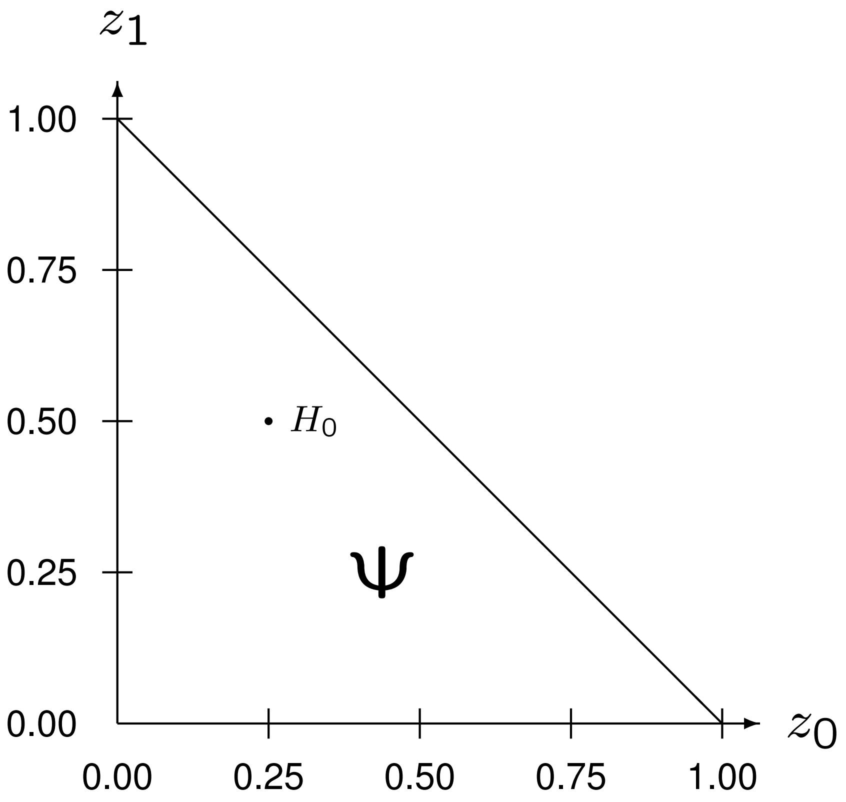
Assume a sample of n nuclear families (each family consisting of two affected sibs and their parents), in which all family members are typed at a marker locus. Further, assume that the marker locus is completely informative. This assumption assures that for each sib pair the number of alleles IBD can be determined unambiguously. For $i = 0, 1, 2$, let n_i denote the observed number of sib pairs sharing i marker alleles IBD $(n_0 + n_1 + n_2 = n)$. Then, (n_2, n_1, n_0) is a realization of a trinomial (i.e., multinomial with $k = 3$, c.f. S/5) distributed random variable (N_2, N_1, N_0) with parameters n and (z_2, z_1, z_0). In case of no linkage, $(z_2, z_1, z_0) = (1/4, 1/2, 1/4)$. Therefore, an ASP test has to decide between the hypotheses

$$H_0 : (z_2, z_1, z_0) = (1/4, 1/2, 1/4) \text{ vs. } H_1 : (z_2, z_1, z_0) \neq (1/4, 1/2, 1/4).$$

Parameter space Ψ for ASP tests

$\psi \hat{=} (z_2, z_1, z_0)$

$\Psi \hat{=} \{(z_2, z_1, z_0) : z_i \geq 0, \sum_{i=0}^{2} z_i = 1\}$



ASP tests: Likelihood ratio test

The maximum likelihood estimate $(\widehat{z}_2, \widehat{z}_1, \widehat{z}_0)$ of (z_2, z_1, z_0) is given by (c.f. S/5)

$$\widehat{z}_i = \frac{n_i}{n}, i = 0, 1, 2.$$

Therefore, the test statistic of the likelihood ratio test (c.f. S/19) is

$$T(n_2, n_1, n_0) = -2 \ln \frac{(1/4)^{n_2} \cdot (1/2)^{n_1} \cdot (1/4)^{n_0}}{(n_2/n)^{n_2} \cdot (n_1/n)^{n_1} \cdot (n_0/n)^{n_0}}.$$

The null distribution of this test statistic can be approximated by the χ^2_2 distribution.

Genetic constraints for IBD distributions

Exercise:

1. (simple) Use the equations given on NPL/6 to show that the ibd probabilities z_2, z_1, and z_0 at the disease locus always satisfy

$$z_1 \leq 1/2 \qquad \text{and} \qquad 2z_0 \leq z_1.$$

2. (more difficult) Use the equations given on NPL/7 to show that the ibd probabilities z_2^M, z_1^M, and z_0^M at the marker locus always satisfy

$$z_1^M \leq 1/2 \qquad \text{and} \qquad 2z_0^M \leq z_1^M.$$

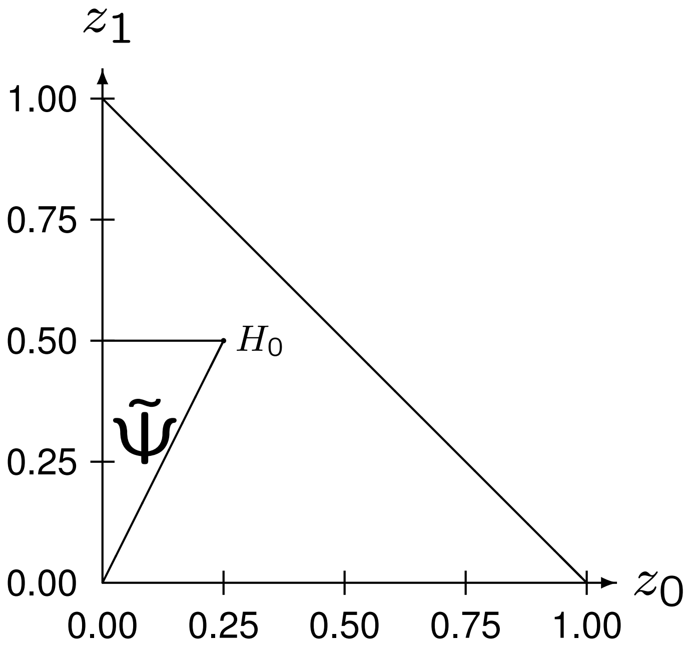
3. (simple) The book by Almgren et al. considers the constraint $3z_1 + 2z_2 \geq 2$. Show that this constraint is equivalent to $2z_0 \leq z_1$, i.e.,

$$3z_1 + 2z_2 \geq 2 \Leftrightarrow 2z_0 \leq z_1.$$

ASP tests obeying the genetic constraints

It can be shown that the constraints $z_1 \leq 1/2$ and $2z_0 \leq z_1$ are satisfied for a broad class of disease models (i.e., not only for single locus disease models).

$\tilde{\Psi} \hat{=} \{(z_2, z_1, z_0) : z_i \geq 0, \sum_{i=0}^{2} z_i = 1, z_1 \leq 1/2, 2z_0 \leq z_1\}$



ASP test: restricted likelihood ratio test

When maximization is restricted to $\tilde{\Psi}$, then the maximum likelihood estimate $(\tilde{z}_2, \tilde{z}_1, \tilde{z}_0)$ of (z_2, z_1, z_0) is given by

$$(\tilde{z}_2, \tilde{z}_1, \tilde{z}_0) = \begin{cases} \left(\frac{n_2}{2(n_2+n_0)}, \frac{1}{2}, \frac{n_0}{2(n_2+n_0)}\right) & \text{for } \frac{n_1}{n} > \frac{1}{2} \text{ and } n_2 > n_0 \\ \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) & \text{for } \frac{n_1}{n} > \frac{1}{2} \text{ and } n_2 \leq n_0 \\ \left(\frac{n_2}{n}, \frac{2(n_1+n_0)}{3n}, \frac{n_1+n_0}{3n}\right) & \text{for } 2\frac{n_0}{n} > \frac{n_1}{n} \text{ and } \frac{n_2}{n} > \frac{1}{4} \\ \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) & \text{for } 2\frac{n_0}{n} > \frac{n_1}{n} \text{ and } \frac{n_2}{n} \leq \frac{1}{4} \\ \left(\frac{n_2}{n}, \frac{n_1}{n}, \frac{n_0}{n}\right) & \text{otherwise} \end{cases}$$

ASP test: restricted likelihood ratio test

The test statistic of the restricted likelihood ratio test ("possible triangle test") is

$$\tilde{T}(n_2, n_1, n_0) = -2 \ln \frac{(1/4)^{n_2} \cdot (1/2)^{n_1} \cdot (1/4)^{n_0}}{(\tilde{z}_2)^{n_2} \cdot (\tilde{z}_1)^{n_1} \cdot (\tilde{z}_0)^{n_0}}.$$

The null distribution of this test statistic can be approximated by a $\left(\frac{1}{2} - \frac{\arccos\sqrt{2/3}}{2\pi}\right) : \frac{1}{2} : \frac{\arccos\sqrt{2/3}}{2\pi}$ mixture of χ^2 distributions with 0, 1, and 2 degrees of freedom.

Maximum lod score (MLS)

The statistic

$$T^{\star}(n_2, n_1, n_0) = -\log \frac{(1/4)^{n_2} \cdot (1/2)^{n_1} \cdot (1/4)^{n_0}}{(\tilde{z}_2)^{n_2} \cdot (\tilde{z}_1)^{n_1} \cdot (\tilde{z}_0)^{n_0}}.$$

is called the *maximum lod score* (MLS) statistic.

Note that

1. $\tilde{T} = 2 \cdot \ln(10) \cdot T^{\star}$
2. Although $T^{\star}$ is named “maximum lod score”, it is not the same as a maximum lod score $Z(\widehat{\theta})$ in parametric linkage analysis (i.e., $T^{\star}$ and $Z(\widehat{\theta})$ possess different null distributions).

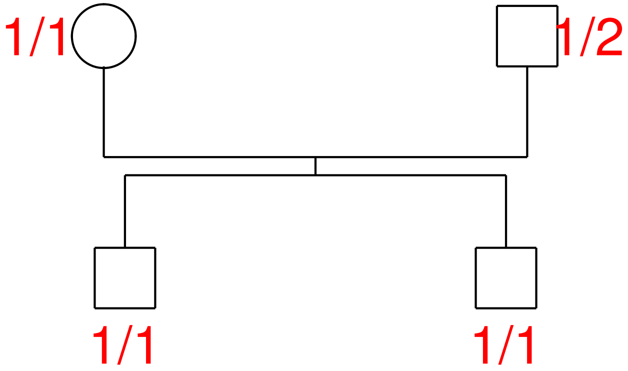
MLS and incompletely informative families

Let PM_j and CM_j denote the observed marker data in the parents and in the children of family j. Let $\mathrm{M}_j = (\mathrm{PM}_j, \mathrm{CM}_j)$ denote the observed marker data in family j. Finally, let IBD_j denote the true (not necessarily observable) number of alleles shared ibd by the children of family j. Then,

$$\begin{aligned}
P(\mathrm{M}_j) &= \sum_{i=0}^{2} P(\mathrm{M}_j \cap (\mathrm{IBD}_j = i)) \\
&= \sum_{i=0}^{2} P(\mathrm{PM}_j \cap \mathrm{CM}_j \cap (\mathrm{IBD}_j = i)) \\
&= \sum_{i=0}^{2} \underbrace{P(\mathrm{IBD}_j = i)}_{z_i} \cdot \underbrace{P(\mathrm{PM}_j \mid \mathrm{IBD}_j = i)}_{P(\mathrm{PM}_j)} \cdot P(\mathrm{CM}_j \mid \mathrm{PM}_j \cap (\mathrm{IBD}_j = i)) \\
&= P(\mathrm{PM}_j) \cdot \sum_{i=0}^{2} z_i \cdot \underbrace{P(\mathrm{CM}_j \mid \mathrm{PM}_j \cap (\mathrm{IBD}_j = i))}_{=: w_{ij}}
\end{aligned}$$

MLS and incompletely informative families

Example:



$$P(\mathrm{CM}_j \mid \mathrm{PM}_j \cap (\mathrm{IBD}_j = 2)) = \frac{1}{2}$$

$$P(\mathrm{CM}_j \mid \mathrm{PM}_j \cap (\mathrm{IBD}_j = 1)) = \frac{1}{4}$$

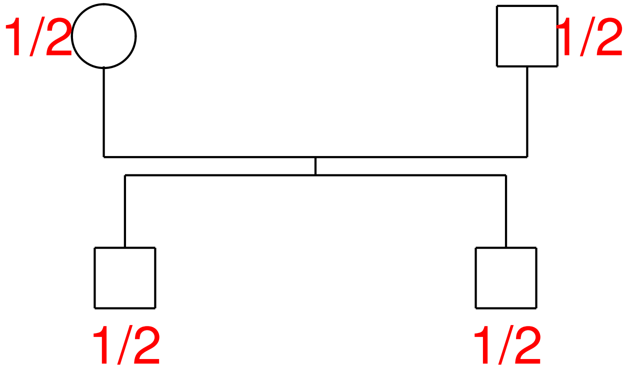
$$P(\mathrm{CM}_j \mid \mathrm{PM}_j \cap (\mathrm{IBD}_j = 0)) = 0$$

$\Rightarrow$ Contribution of this family to the likelihood is

$$\frac{1}{2}\, z_2 + \frac{1}{4}\, z_1$$

MLS and incompletely informative families

Example:



$$P(\mathrm{CM}_j \mid \mathrm{PM}_j \cap (\mathrm{IBD}_j = 2)) = \frac{1}{2}$$

$$P(\mathrm{CM}_j \mid \mathrm{PM}_j \cap (\mathrm{IBD}_j = 1)) = 0$$

$$P(\mathrm{CM}_j \mid \mathrm{PM}_j \cap (\mathrm{IBD}_j = 0)) = \frac{1}{2}$$

$\Rightarrow$ Contribution of this family to the likelihood is

$$\frac{1}{2}\, z_2 + \frac{1}{2}\, z_0$$

MLS and incompletely informative families

Likelihood of the whole sample x:

$$L(z_2, z_1, z_0 \mid x) = \prod_{j=1}^{n} \left(\sum_{i=0}^{2} z_i \cdot w_{ij} \right)$$

- How to obtain the restricted ML-estimates for (z_2, z_1, z_0)?

  $\rightarrow$ EM-algorithm

- Null distribution of the restricted likelihood ratio test?

  This distribution can still be approximated by a mixture of χ^2_n

  $(n = 0, 1, 2)$ distributions.

Relationship between MLS and parametric linkage analysis

Nonparametric methods of linkage analysis are motivated by the difficulty to specify an appropriate disease model, which is required for parametric linkage analysis. On PL/25, MOD score analysis (i.e., calculation of the lod score maximized not only over θ, but also over the disease model parameters (f_2, f_1, f_0, p_T)) was mentioned to circumvent this problem of disease model specification. It can be shown that for samples of affected sib pairs, MOD score analysis and MLS are identical.