Sibling Pair Linkage Tests

Biostatistics 666

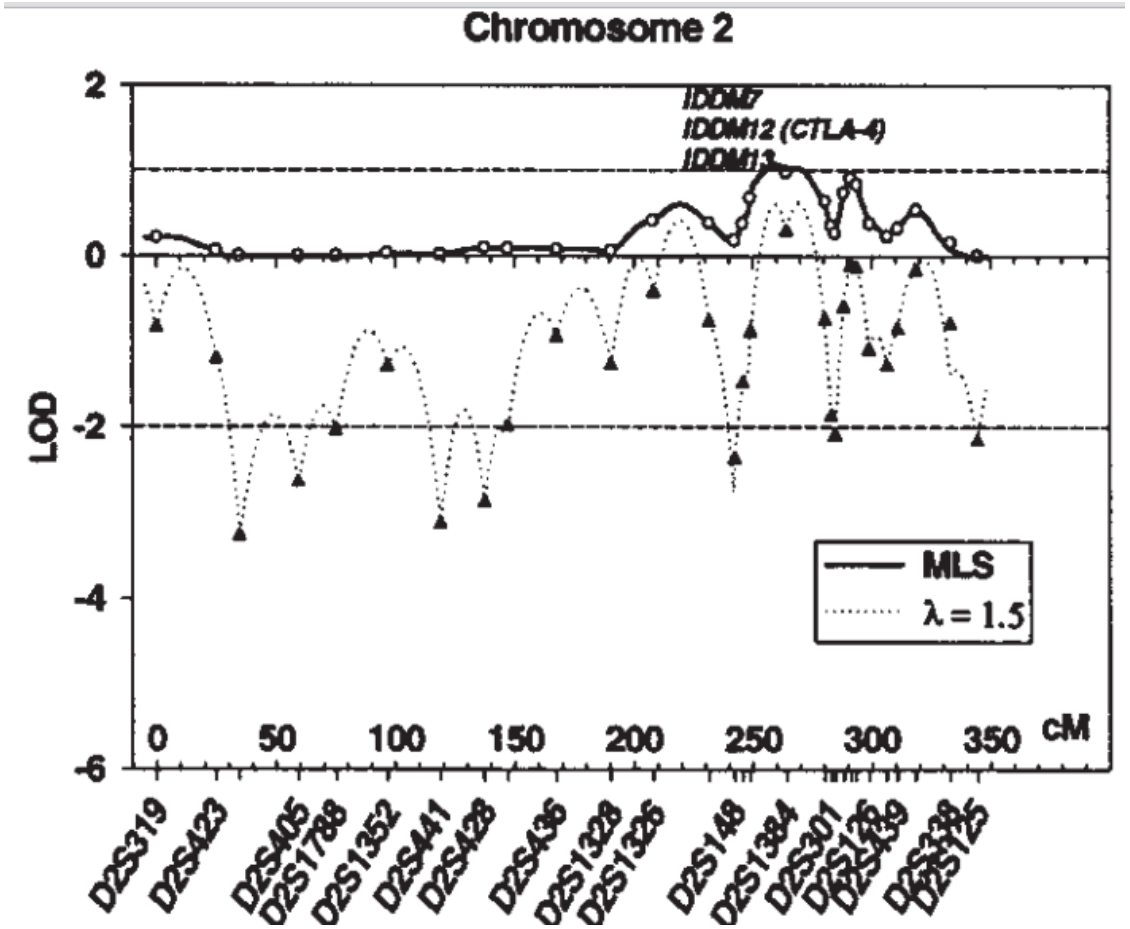
Today …

- Introduction to linkage analysis of affected siblings
- A simple disease model
  - Probability of sampling affected relative pairs
- Linkage analysis with sibling pairs using Risch's Maximum LOD Score (MLS)
- Distribution of IBD in affected sibling pairs and Holman's "Possible Triangle Constraint"

Examplar Linkage Study

- Concannon et al (1998) *Nature Genetics,* **19:**292-296
- Affected sibling pair study of type 1 diabetes
  - Common chronic disease of childhood
  - 292 affected sibpairs for initial screen
  - 467 affected sibpairs for follow-up
- Highest LOD score reaches 34.2 near HLA on chr. 6
  - At this locus, chromosomes carried by affected siblings are identical 73% of the time.

Examplar Linkage Study Results



Concannon et al (1998) *Nature Genetics*, **19:**292-296

Single Locus Disease Model

1. Allele frequencies
   - For normal and susceptibility alleles
2. Penetrances
   - Probability of disease for each genotype

- Useful in exploring behavior of linkage and association tests
- Simplification of reality, ignores other loci and the environment

Penetrance

- $f_{ij} = P(\ Affected \mid G = ij)$
- Probability someone with genotype *ij* is affected
- Models the marginal effect of each locus

Using Penetrances

- Allele frequency p
- Genotype penetrances f_{11}, f_{12}, f_{22}

- Probability of genotype given disease
  - P(G = ij | D) =
- Prevalence
  - K =

Pairs of Individuals

- A genetic model can predict probability of sampling different affected relative pairs
- We will consider some simple cases:
  - Unrelated individuals
  - Parent-offspring pairs
  - Monozygotic twins
- What do the pairs above have in common?
  - HINT: Think about the amount of shared genetic material

What we might expect …

- Related individuals have similar genotypes
- For a genetic disease…
- Probability that two relatives are both affected must be greater or equal to the probability that two randomly sampled unrelated individuals are affected

Relative Risk and Prevalence

- In relation to affected proband, define
  - K_R prevalence in relatives of type R
  - $\lambda_R = K_R/K$ increase in risk for relatives of type R
- λ_R is a measure of the overall effect of a locus
  - Useful for predicting power of linkage studies

Unrelated Individuals

- Probability of affected pair of unrelateds

$$
\begin{aligned}
P(a \text{ and } b \text{ affected}) &= \mathrm{P}(a \text{ affected})\mathrm{P}(b \text{ affected}) \\
&= \mathrm{P}(\text{affected})^2 \\
&= \left[p^2 f_{11} + 2p(1-p) f_{12} + (1-p)^2 f_{22}\right]^2 \\
&= K^2
\end{aligned}
$$

- For any two related individuals, probability that both are affected should be greater

Monozygotic Twins

- Probability of affected pair of identical twins

$$
\begin{aligned}
P(MZ \text{ pair affected}) &= \sum_G P(G)P(a \text{ affected} \mid \mathrm{G})P(b \text{ affected} \mid G) \\
&= p^2 f_{11}^2 + 2p(1-p) f_{12}^2 + (1-p)^2 f_{22}^2 \\
&= K_{MZ} K \\
&= \lambda_{MZ} KK
\end{aligned}
$$

- λ_{MZ} will be greater than for any other relationship

Parent Offspring Pairs

- Probability of affected parent-offspring pair

$$
\begin{aligned}
P &= P(\,parent \text{ and } child \text{ affected}) \\
&= \sum_{G_P} \sum_{G_O} P(G_P, G_O)\, f_{G_P}\, f_{G_O} \\
&= \sum_{i} \sum_{j} \sum_{k} P(i, j, k)\, f_{ij}\, f_{ik} \\
&= p^3 f_{11}^2 + (1-p)^3 f_{22}^2 + p(1-p) f_{12}^2 + 2p^2(1-p) f_{11} f_{12} + 2p(1-p)^2 f_{22} f_{12} \\
&= KK_O \\
&= \lambda_O KK
\end{aligned}
$$

- λ_o will be between 1.0 and λ_{MZ}

IBD – Identity by Descent

- Sharing of segregating stretch of chromosome within a family
- If a stretch of chromosome is shared IBD, all variants within the stretch will be shared
- At any locus siblings share 0, 1 or 2 alleles IBD
  - Baseline probabilities of IBD 0, 1 and 2 are ¼, ½ and ¼

For a single locus model...

$$\lambda_{IBD=2} = \lambda_{MZ}$$
$$\lambda_{IBD=1} = \lambda_{O}$$
$$\lambda_{IBD=0} = 1$$

$$K_{IBD=2} = K_{MZ}$$
$$K_{IBD=1} = K_{O}$$
$$K_{IBD=0} = K$$

- Model ignores contribution of other genes and environment
- Simple model that allows for useful predictions
  - Risk to half-siblings
  - Risk to cousins
  - Risk to siblings

Point of Situation

- Probabilities of affected pairs for
  - Unrelated Individuals
  - Monozygotic Twins
  - Parent-Offspring Pairs
- Each of these shares a fixed number of alleles IBD …

Affected Half-Siblings

- IBD sharing
  - 0 alleles with probability 50%
  - 1 allele with probability 50%

- This gives …

$$\lambda_H = \tfrac{1}{2}\lambda_O + \tfrac{1}{2} = \tfrac{1}{2}(\lambda_O + 1)$$

$$K_H = \tfrac{1}{2}K_O + \tfrac{1}{2}K = \tfrac{1}{2}(K_O + K)$$

Affected Sibpairs

- IBD sharing …
  - 0 alleles with probability 25%
  - 1 alleles with probability 50%
  - 2 alleles with probability 25%

- This gives …

$$\lambda_S = \tfrac{1}{4}\lambda_{MZ} + \tfrac{1}{2}\lambda_O + \tfrac{1}{4} = \tfrac{1}{4}(\lambda_{MZ} + 2\lambda_O + 1)$$

which implies

$$\lambda_{MZ} = 4\lambda_S - 2\lambda_O - 1$$

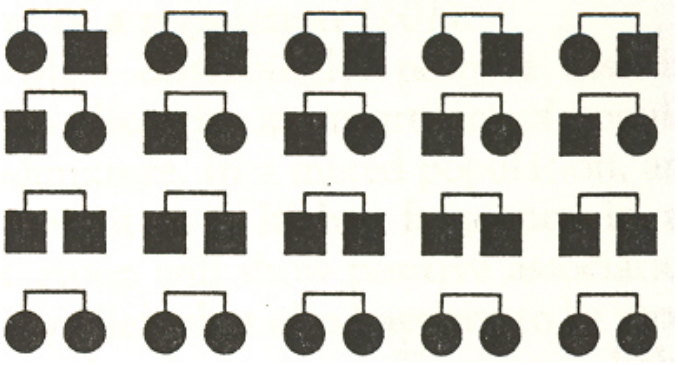
Important Notes…

- We can use allele frequencies and penetrances to estimate probability of affected relative pairs

- Among sibling pairs, pairs with two alleles “identical-by-descent” have the highest probability of both being affected
  - Most like “identical twins” for single locus models

Affected Sibpair Linkage Analyses

- Consider affected sibling pairs
- Consider one genetic marker at a time
- Are paired genotypes more similar than expected?
- Only a subset of all genetic markers must be examined



Likelihood Based Linkage Test

- Depends on three parameters z_0, z_1, z_2
  - Probability of sharing 0, 1 and 2 alleles IBD
- Null likelihood uses $z_0=¼$, $z_1=½$, $z_2=¼$
- Alternative likelihood uses MLE for z_0, z_1, z_2
- Compare likelihoods with likelihood ratio test

Potential Sib-Pair Likelihood

Under the null hypothesis:

$$L = (1/4)^{n_{IBD0}} (1/2)^{n_{IBD1}} (1/4)^{n_{IBD2}}$$

Under the alternative hypothesis

$$L = (\hat{z}_0)^{n_{IBD0}} (\hat{z}_1)^{n_{IBD1}} (\hat{z}_2)^{n_{IBD2}}$$

Likelihood Ratio Based Test Statistics

$$LOD = \log_{10} \frac{L(\hat{z}_0, \hat{z}_1, \hat{z}_2)}{L(z_0 = \tfrac{1}{4}, z_1 = \tfrac{1}{2}, z_2 = \tfrac{1}{4})}$$

$$\chi^2 = 2 \ln \frac{L(\hat{z}_0, \hat{z}_1, \hat{z}_2)}{L(z_0 = \tfrac{1}{4}, z_1 = \tfrac{1}{2}, z_2 = \tfrac{1}{4})}$$

$$= 2 \ln L(\hat{z}_0, \hat{z}_1, \hat{z}_2) - 2 \ln L(z_0 = \tfrac{1}{4}, z_1 = \tfrac{1}{2}, z_2 = \tfrac{1}{4})$$

In real life…

- Markers are only partially informative
- IBD sharing is equivocal
  - Uncertainty can only be partly reduced by examining relatives
- Need an alternative likelihood
  - Should allow for partially informative data

Desirable Properties

- Models IBD probabilities z_0, z_1, z_2
  - Probability of sharing 0, 1 and 2 alleles IBD
- Uses partial information on IBD sharing
- For unambiguous data, equivalent to previous likelihood

For A Single Family

$$L_i = \sum_{j=0}^{2} P(IBD = j \mid ASP) P(Genotypes_i \mid IBD = j) = \sum_{j=0}^{2} z_j w_{ij}$$

Risch (1990) defines

$$w_{ij} = P(Genotypes_i \mid IBD = j)$$

We only need proportionate w_{ij}

Likelihood and LOD Score

$$L(z_0, z_1, z_2) = \prod_i \sum_j z_j w_{ij}$$

$$LOD = \log_{10} \prod_i \frac{\hat{z}_0 w_{i0} + \hat{z}_1 w_{i1} + \hat{z}_2 w_{i2}}{\frac{1}{4} w_{i0} + \frac{1}{2} w_{i1} + \frac{1}{4} w_{i2}}$$

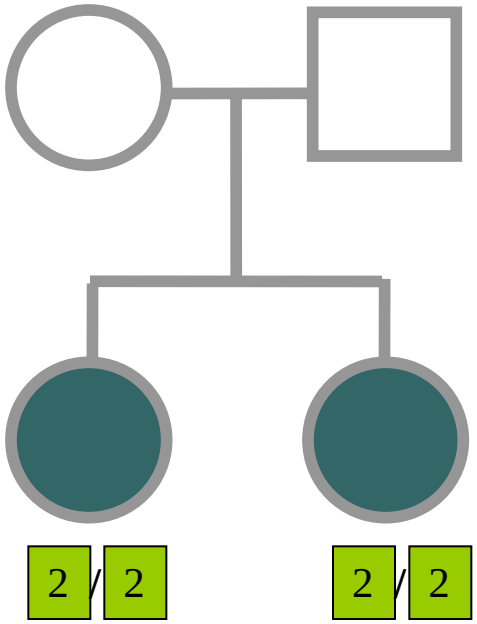
The MLS statistic is the LOD evaluated at the MLEs of z_0, z_1, z_2

P(Marker Genotype|IBD State)

Relative		IBD		
I	**II**	0	1	2
(a,b)	(c,d)	$4p_ap_bp_cp_d$	0	0
(a,a)	(b,c)	$2p_a^2p_bp_c$	0	0
(a,a)	(b,b)	$p_a^2p_b^2$	0	0
(a,b)	(a,c)	$4p_a^2p_bp_c$	$p_ap_bp_c$	0
(a,a)	(a,b)	$2p_a^3p_b$	$p_a^2p_b$	0
(a,b)	(a,b)	$4p_a^2p_b^2$	$(p_ap_b^2+p_a^2p_b)$	$2p_ap_b$
(a,a)	(a,a)	p_a^4	p_a^3	p_a^2
Prior Probability		¼	½	¼

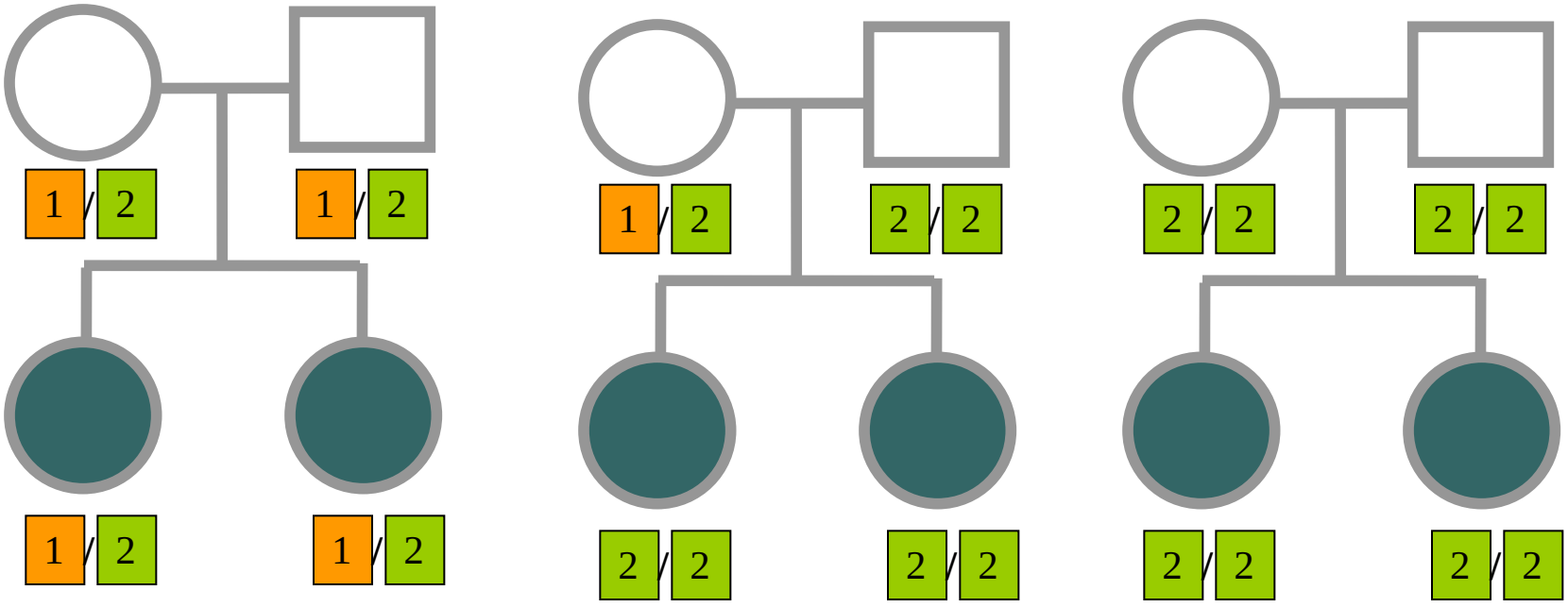
These probabilities apply to pair of individuals, when no other genotypes in the family are known.

Example scoring for w_{ij}



In this case, relative weights depend on allele frequency.

More examples for scoring: w_{ij}



In these cases, multiple weights are non-zero (but equal) for each family.

How to maximize likelihood?

- If all families are informative
  - Use sample proportions of IBD=0, 1, 2
- If some families are uninformative
  - Use an E-M algorithm
  - At each stage generate complete dataset with fractional counts
  - Iterate until estimates of LOD and z parameters are stable

Assigning Partial Counts in E-M

$$P(IBD = j \mid Genotypes) =$$

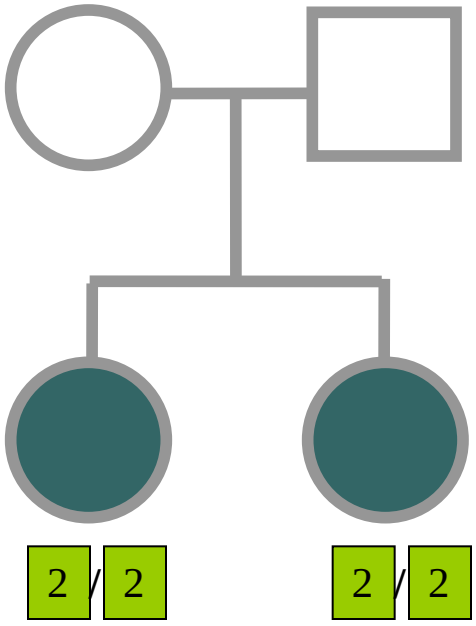
$$= \frac{P(IBD = j \mid ASP)P(Genotypes \mid IBD = j)}{L_i}$$

$$= \frac{P(IBD = j \mid ASP)P(Genotypes \mid IBD = j)}{\sum_{k=0}^{2} P(IBD = k \mid ASP)P(Genotypes \mid IBD = k)}$$

$$= \frac{z_j w_{ij}}{\sum_{k=0}^{2} z_k w_{ik}}$$

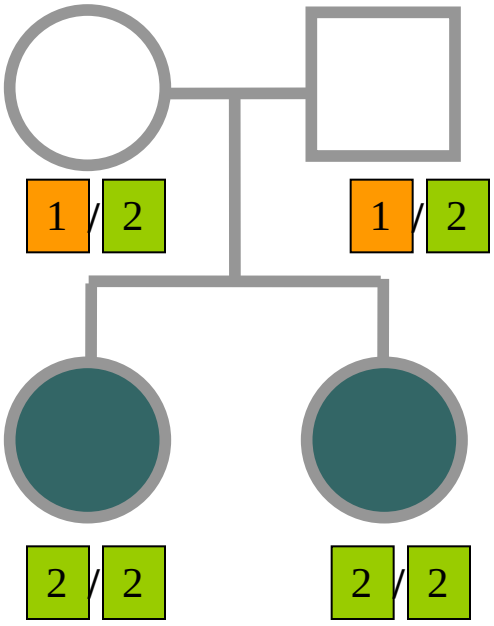
Example

5x



IBD=?

5x



IBD=2

Assume a bi-allelic marker where the two alleles have identical frequencies.

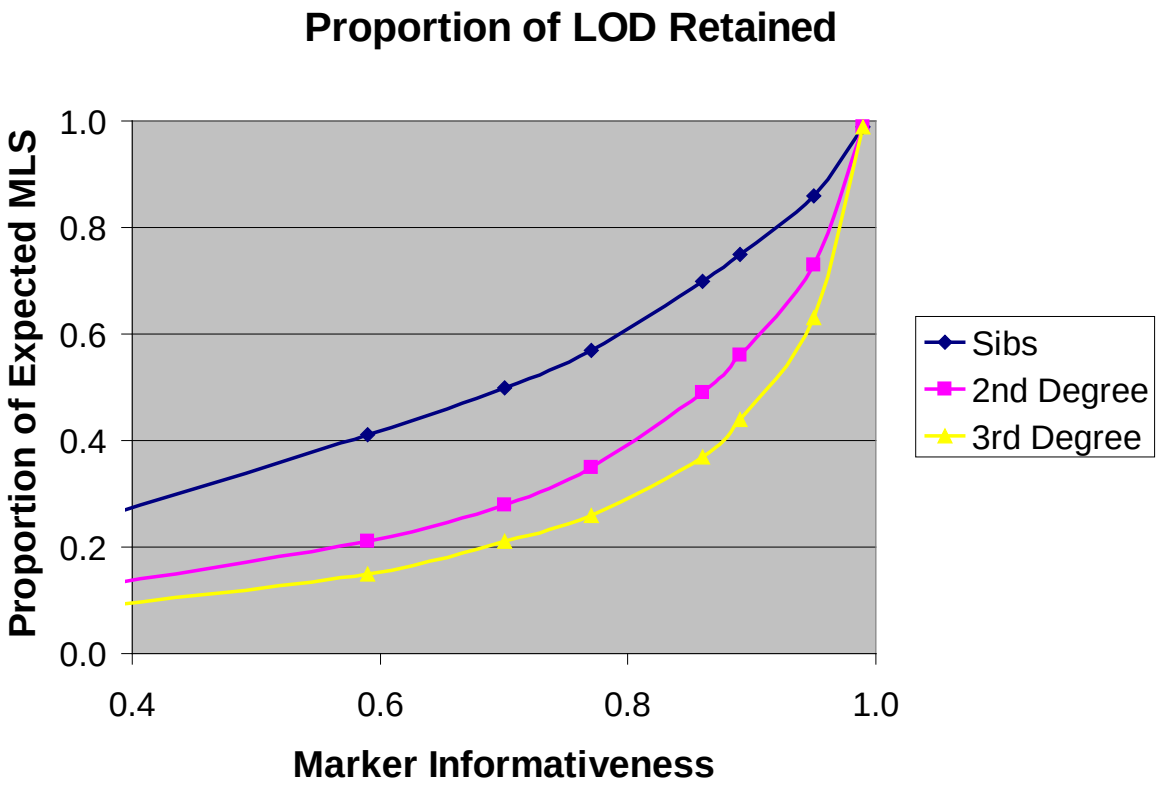
Example of E-M Steps

Parameters			Equivocal Families			Other			
z0	**z1**	**z2**	**IBD=0**	**IBD=1**	**IBD=2**	**IBD=2**	**LOD**	**LODi**	**LODu**
0.250	0.500	0.250	0.56	2.22	2.22	5	0.00	0.00	0.00
0.056	0.222	0.722	0.08	0.66	4.26	5	3.19	2.30	0.89
0.008	0.066	0.926	0.01	0.17	4.82	5	4.01	2.84	1.16
0.001	0.017	0.982	0.00	0.04	4.96	5	4.20	2.97	1.23
0.000	0.004	0.996	0.00	0.01	4.99	5	4.25	3.00	1.24
0.000	0.001	0.999	0.00	0.00	5.00	5	4.26	3.01	1.25
0.000	0.000	1.000	0.00	0.00	5.00	5	4.26	3.01	1.25

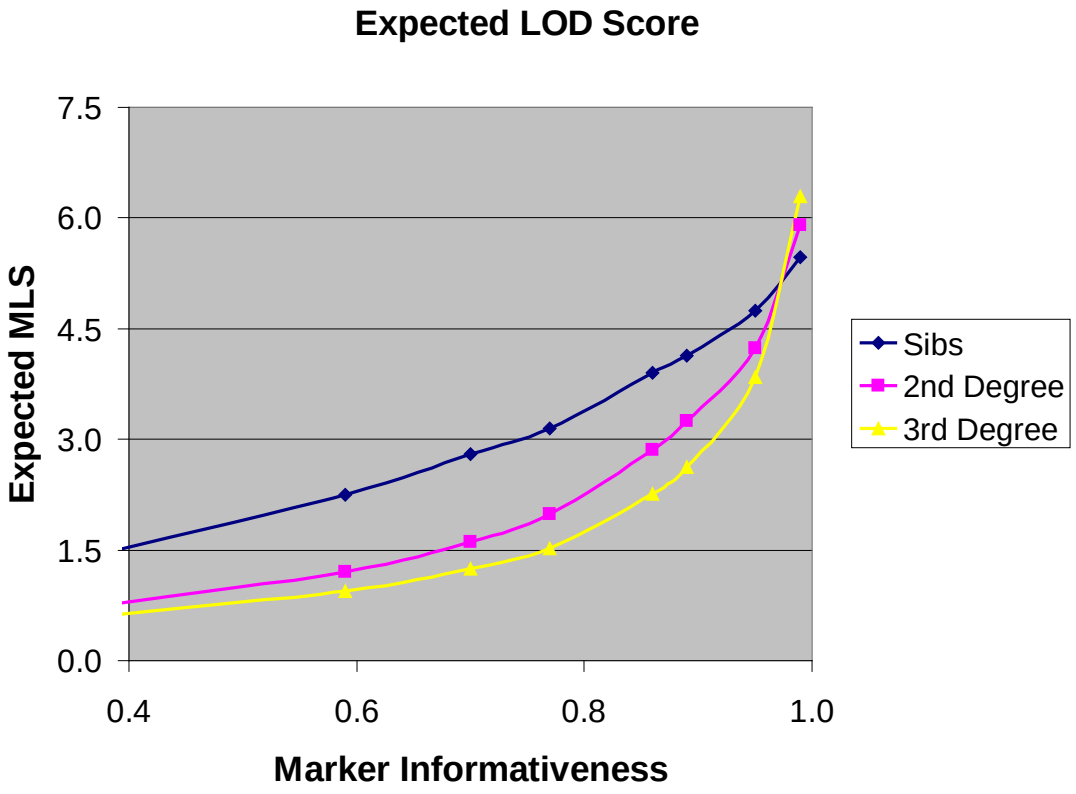
Properties of Pair Analyses Explored by Risch

- Effect of marker informativeness
- Effect of adding relative genotypes
- Size of genetic effect
- Degree of relationship

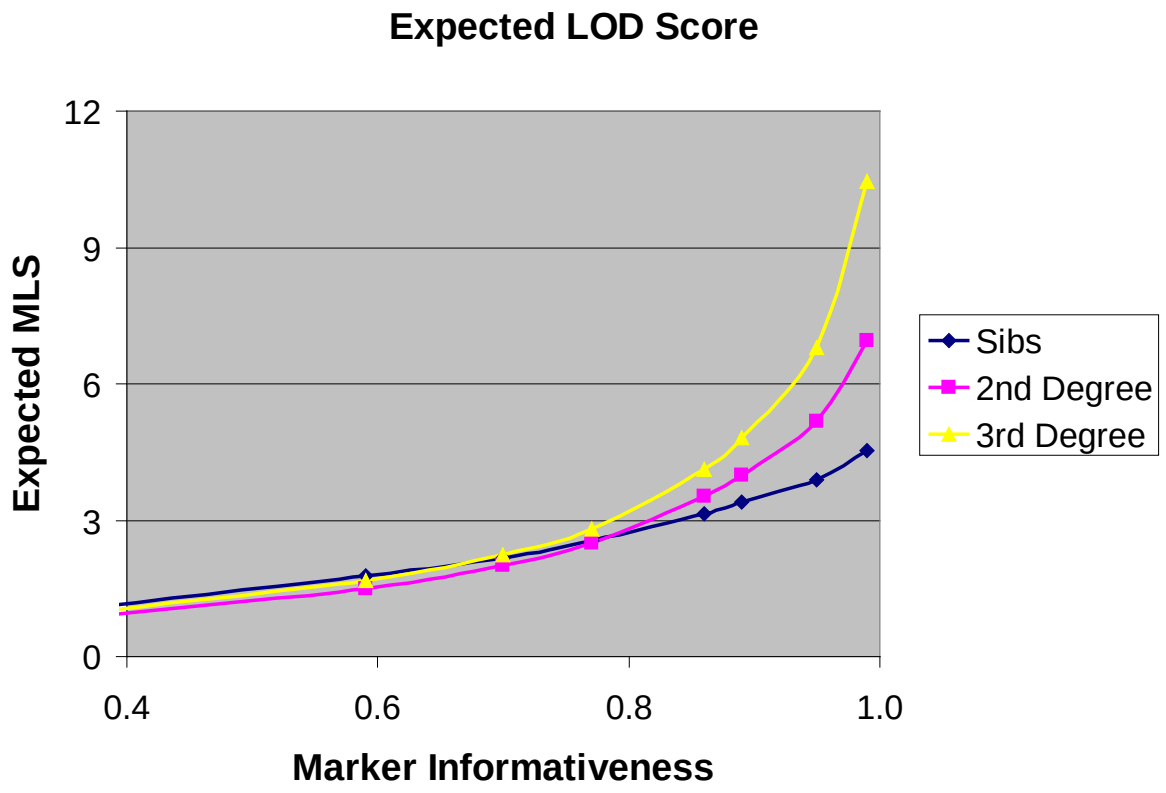
Marker Informativeness



Marker Informativeness
Gene of Modest Effect (λ_O=3)



Marker Informativeness
Gene of Larger Effect (λ_O=10)



Genotypes of Other Family Members

- Genotyping only pair decreases LOD score by
  - Up to 33% if only sib-pairs are genotyped
  - Up to 60% for second degree relatives
  - Up to 70% for third degree relatives

- Genotyping effort decreases by
  - 50% if only sib-pairs are typed
  - 60% if only second degree relatives typed
  - 75% if only third degree relatives typed

Point of Situation …

- Noted that affected siblings are more likely to share two alleles identical by descent
- Derived a likelihood based linkage test that compares sharing probabilities to null defaults
- Let’s examine these probabilities in more detail …

Next ...

- Predicting distribution of IBD
  - Modeling marginal effect of a single locus
  - Relative risk ratio (λ_R)
- The Possible Triangle for Sibling Pairs
  - Plausible IBD values for affected siblings
  - Refinement of the model of Risch (1990)

Recurrence Risks vs. IBD

$$\lambda_{IBD=2} = \lambda_{MZ} = \frac{P(affected \mid IBD = 2 \text{ with affected relative})}{P(affected)}$$

$$\lambda_{IBD=1} = \lambda_{O} = \frac{P(affected \mid IBD = 1 \text{ with affected relative})}{P(affected)}$$

$$\lambda_{IBD=0} = 1 = \frac{P(affected \mid IBD = 0 \text{ with affected relative})}{P(affected)}$$

Bayes' Theorem: Predicting IBD Sharing

$$P(IBD = i \mid \text{affected pair}) =$$

$$= \frac{P(IBD = i)P(\text{affected pair} \mid IBD = i)}{\sum_j P(IBD = j)P(\text{affected pair} \mid IBD = j)}$$

$$= \frac{P(IBD = i)\lambda_{IBD=i}}{\sum_j P(IBD = j)\lambda_{IBD=j}}$$

Sibpairs
Expected Values for z_0, z_1, z_2

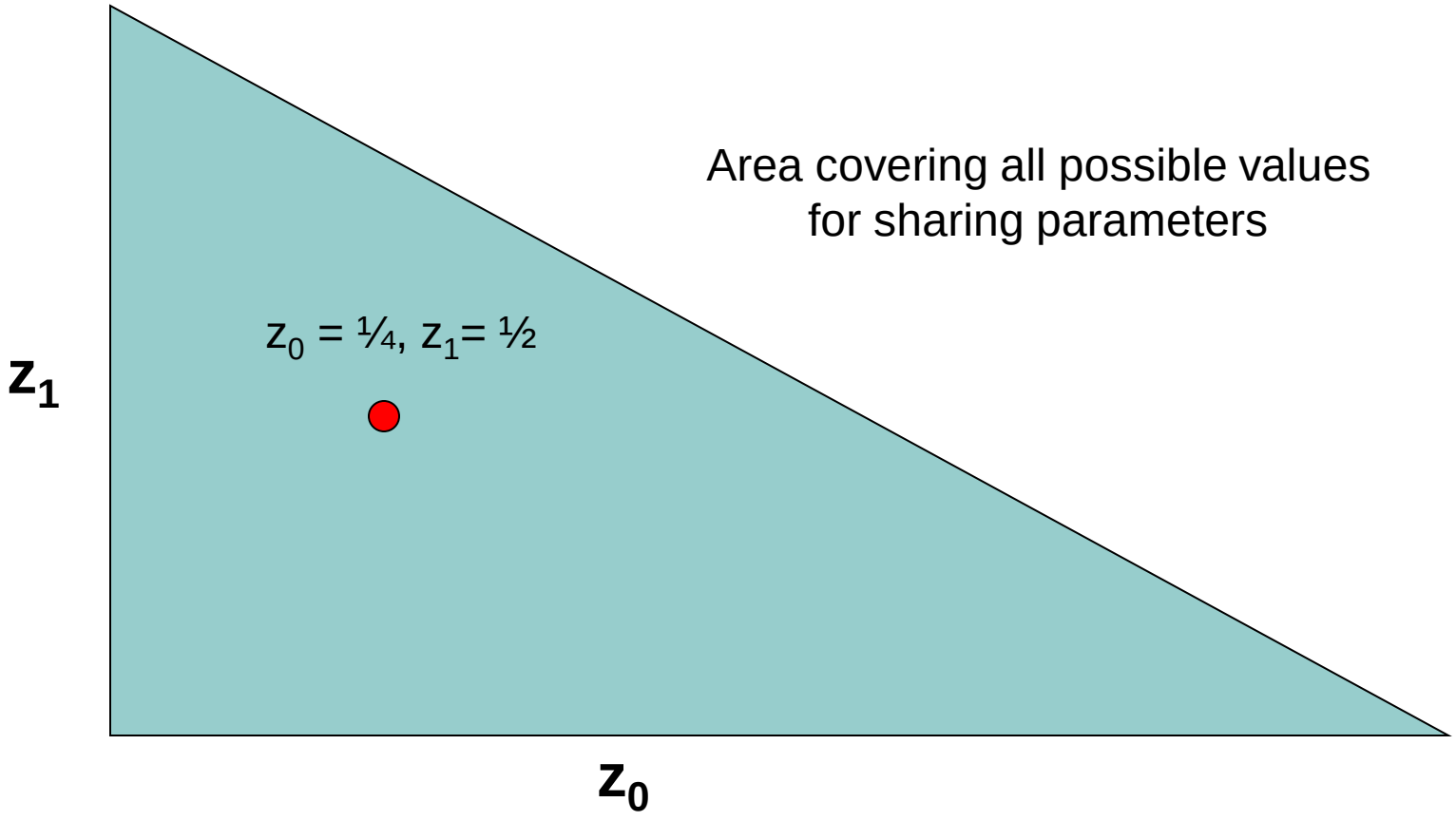
$$z_0 = 0.25\frac{1}{\lambda_s}$$

$$z_1 = 0.50\frac{\lambda_o}{\lambda_s}$$

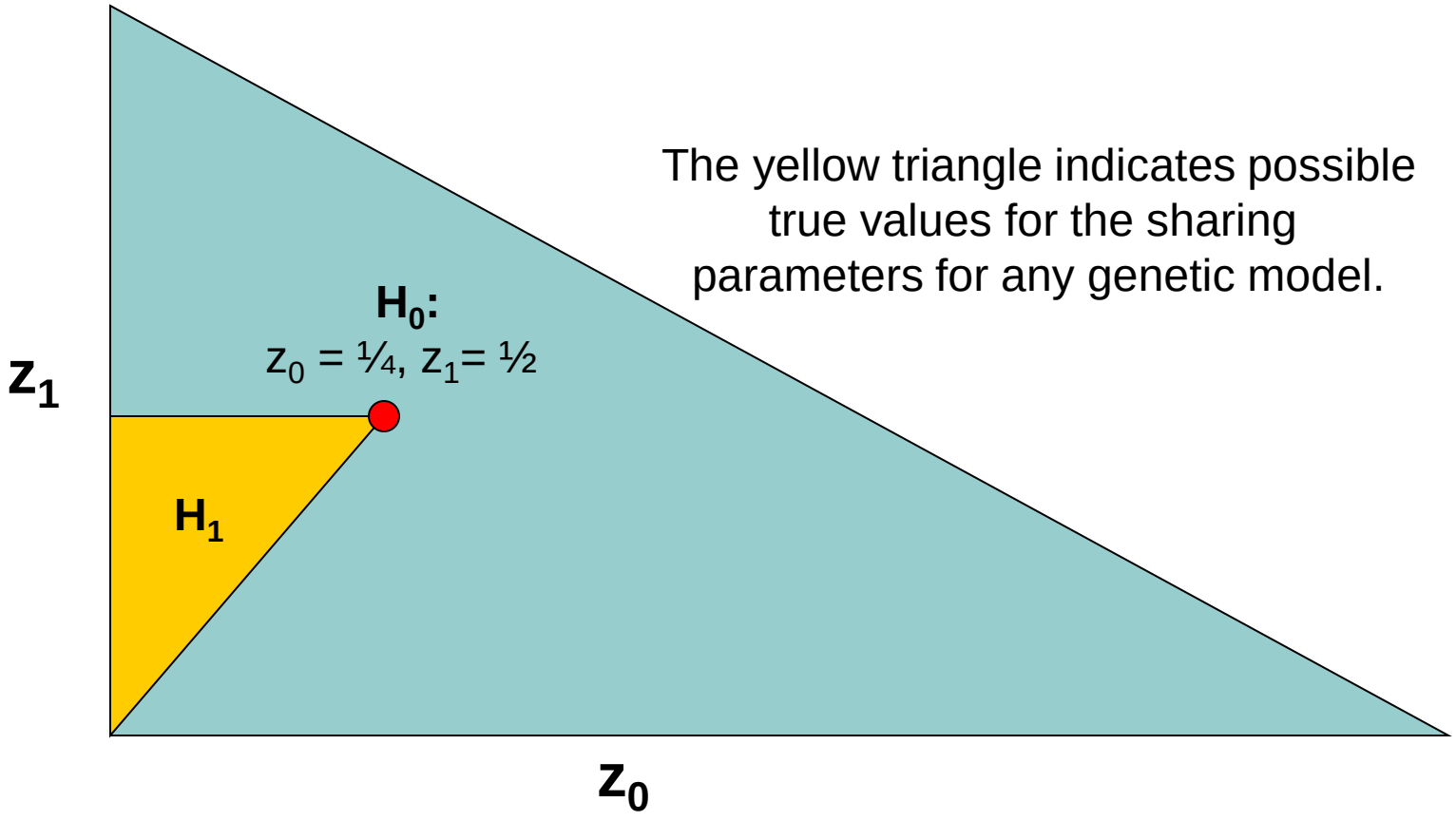
$$z_2 = 0.25\frac{\lambda_{MZ}}{\lambda_s}$$

$1 \le \lambda_o \le \lambda_s \le \lambda_{MZ}$ for any genetic model

Possible Triangle



Possible Triangle



The yellow triangle indicates possible true values for the sharing parameters for any genetic model.

Intuition

- Under the null
  - True parameter values are (¼, ½, ¼)
  - Estimates will wobble around this point
- Under the alternative
  - True parameter values are within triangle
  - Estimates will wobble around true point

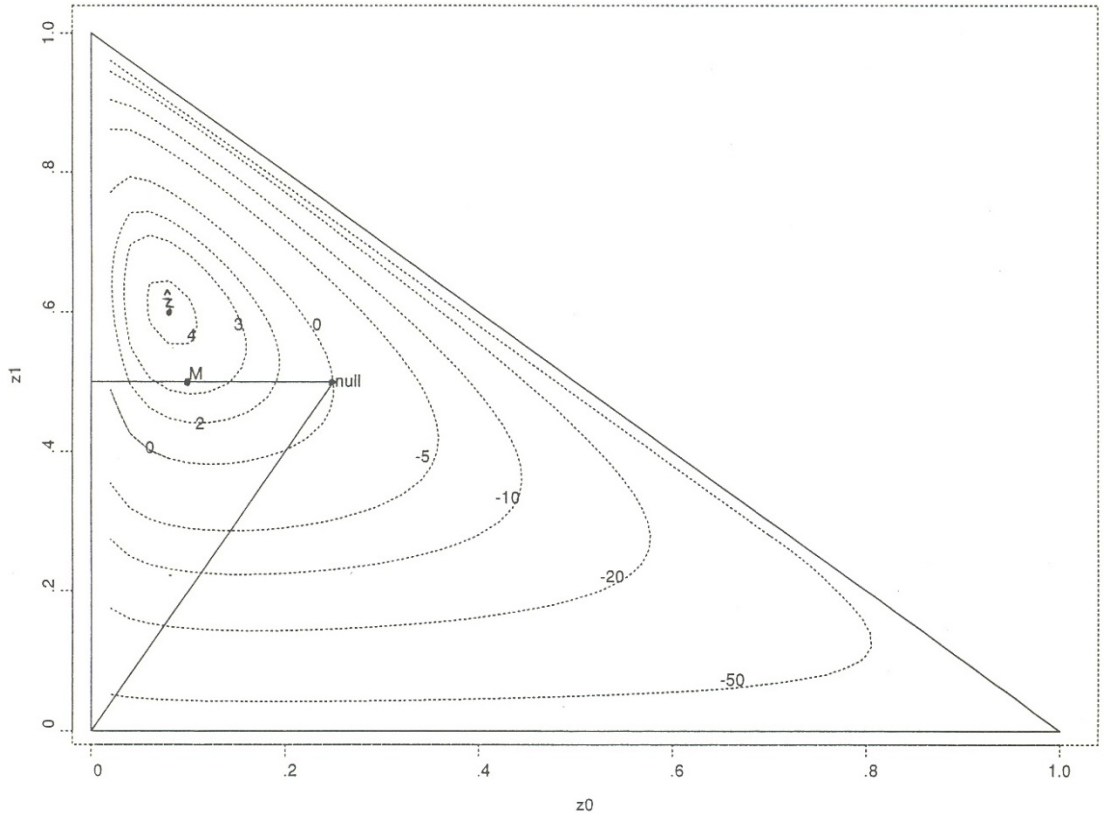
Idea (Holmans, 1993)

- Testing for linkage
  - Do IBD patterns suggest a gene is present?
- Focus on situations where IBD patterns are compatible with a genetic model
  - Restrict maximization to possible triangle

The possible triangle method

1. Estimate z_0, z_1, z_2 without restrictions
2. If estimate of $z_1 > ½$ then …
   a) Repeat estimation with $z_1 = ½$
   b) If this gives $z_0 > ¼$ then revert to null (MLS=0)
3. If estimates imply $2z_0 > z_1$ then …
   a) Repeat estimation with $z_1 = 2z_o$
   b) If this gives $z_0 > ¼$ then revert to null (MLS=0)
4. Otherwise, leave estimates unchanged.

Possible Triangle



Holman's Example:

IBD	Pairs
0	8
1	60
2	32

MLS = 4.22 (overall)
MLE = (0.08,0.60,0.32)

MLS = 3.35 (triangle)
MLE = (0.10,0.50,0.40)

MLS Combined With Possible Triangle

- Under null, true **z** is a corner of the triangle
  - Estimates will often lie outside triangle
  - Restriction to the triangle decreases MLS
  - MLS threshold for fixed type I error decreases
- Under alternative, true **z** is within triangle
  - Estimates will lie outside triangle less often
  - MLS decreases less
  - Overall, power should be increased

Example

- Type I error rate of 0.001
- LOD of 3.0 with unrestricted method
  - Risch (1990)
- LOD of 2.3 with possible triangle constraint
  - Holmans (1993)
  - For some cases, almost doubles power

Recommended Reading

- Holmans (1993)
  Asymptotic Properties of
  Affected-Sib-Pair Linkage Analysis
  Am J Hum Genet **52:**362-374

- Introduces possible triangle constraint
- Good review of MLS method

Recommended Reading

- Risch (1990)
  - Linkage Strategies for Genetically Complex Traits. III. The Effect of Marker Polymorphism on Analysis of Affected Relative Pairs
  - *Am J Hum Genet* **46:**242-253
- Introduces MLS method for linkage analysis
  - Still, one of the best methods for analysis pair data
- Evaluates different sampling strategies
  - Results were later corrected by Risch (1992)